Composite-Based Structural Equation Models

Rainer Schlittgen

January 30, 2019

Introduction

The paper gives the background of the R package `cbsem`. It descibes the computation of covariance matrices of the indicators, the simulation, estimation and segmentation of composite-based structural equation models and gives examples for all methods described.

Two complementary schools have come to the fore in the field of Structural Equation Modelling (SEM): factor-based SEM and composite-based SEM. The first approach has been developed around Karl JÃűreskog and the second one around Herman Wold (Wold 1983, Lohmöller 1989) under the name "PLS" (Partial Least Squares). Hwang and Takane have proposed an other composite-based SEM method named Generalized Structured Component Analysis (Hwang and Takane 2004). Factor-based SEM is usually used with an objective of model validation and needs a large sample. Composite-based SEM is mainly used for score computation and can be carried out on very small samples. Composite-based structural equation models consider linear combinations of the observables or indicators as composites. Through them the relations between blocks of indicators are modeled.

SEMs are visualized with the help of path diagrams. The relations between the variables are shown by arrows pointing to the dependent variables. Composite-based SEMs deal mainly with arrows pointing from the indicators to the composites. These are the weights. The relations are called formative. Sometimes a factor analysis point of view is incorporated, too. Then arrows pointing in the other direction are also present. They are called lodings according to the factor analysis convenience. The corresponding relations are called reflective.

The setting is always given by two blocks of indicators. The correletion structure between this two blocks is modeled by the composites built from the indicators. Three scenarios are dealt with. One has only arrows pointing from the indicators to the composites, in the second all arrows pointing into the other direction are also present and in the third such arrows are present only for one part of indicators.

This paper brings together what I did in the field of composite-based SEMs. I was introduced to this field by a colleague, Prof. Dr. Christian Ringle, who became a good friend. After the first project he fed me with new problems. At the end I was involved in this area over fifteen years. Through the time my point of view evolved to the present one. This is presented here. The most challenging problem was to develop a suitabe method to simulate composite-based SEMs with loadings. Weights and loadings show some interplay which is mostly neglegted.

Contents

1 The GSC model

1.1 The composite-based model

Let two sets of variables be given, $\boldsymbol{x} = (X_1, \ldots, X_{p_1})$ and $\boldsymbol{y} = (Y_1, \ldots, Y_{p_2})$. All the variables should be standardised, $\mathrm{E}(X_i) = 0$ and $\mathrm{Var}(X_i) = 1$, with the same applying to Y_i. The relationships between these two sets of variables are modelled via composites, the linear combinations of the $\boldsymbol{x}$ and $\boldsymbol{y}$ indicator variables. The composites of the $\boldsymbol{x}$ variables are denoted by ξ. These are the exogenous variables that do not depend on any other composite. Each of the composites η, which result from the $\boldsymbol{y}$ indicator variables, is endogenous, depending on at least one other composite, regardless of whether it is a ξ or another η. The number of exogenous composites is q_1, while the number of endogenous composites is q_2.

The observed variables are indicators of their composites. Each composite should have its own set of indicators. The indicators of ξ_g build a subvector $\boldsymbol{x}_g$ of $\boldsymbol{x}$, $g = 1, \ldots, q_1$. The corresponding weights vectors are denoted by $\mathbf{w}_g^{(1)}$. η_h has indicators $\boldsymbol{y}_h$ with wights $\mathbf{w}_h^{(2)}$, $h = 1, \ldots, q_2$. The parameter vectors are column vectors. The random vectors however are, however, row vectors. The weights relations are:

$$\boldsymbol{\xi} = \boldsymbol{x}\mathbf{W}_1\,, \tag{1a}$$

$$\boldsymbol{\eta} = \boldsymbol{y}\mathbf{W}_2\,, \tag{1b}$$

with

$$\mathbf{W}_1 = \begin{pmatrix} \mathbf{w}_1^{(1)} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{w}_2^{(1)} & & \mathbf{0} \\ \vdots & & & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{w}_{q_1}^{(1)} \end{pmatrix}, \quad \mathbf{W}_2 = \begin{pmatrix} \mathbf{w}_1^{(2)} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{w}_2^{(2)} & & \mathbf{0} \\ \vdots & & & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{w}_{q_2}^{(2)} \end{pmatrix}. \tag{1c}$$

Equation (1) formalises the formative part of the model in the structural equation modelling's standard terminology.

The composites should have unit variances, $\mathrm{Var}(\xi_g) = 1$ and $\mathrm{Var}(\eta_h) = 1$. Therefore the weights need to be standardised. They must fulfill $\mathbf{w}_g^{(1)'} \boldsymbol{\Sigma}_{\boldsymbol{x}_g \boldsymbol{x}_g} \mathbf{w}_g^{(1)} = 1$. The same applies to $\mathbf{w}_h^{(2)}$.

The structural model provides the relationships between the two sets of indicators by means of the resulting two sets of composites:

$$\boldsymbol{\eta} = \boldsymbol{\xi}\boldsymbol{\Gamma}' + \boldsymbol{\eta}\mathbf{B}' + \boldsymbol{\zeta}\,, \tag{2}$$

The matrix $\mathbf{B}$ can be arranged as a lower triangular with zeros on the diagonal for recursive models. This should be the case here. $\boldsymbol{\zeta}$ is a vector of errors. The errors are presumed to be uncorrelated and also uncorrelated in respect of the other random vectors present. The formulation with row vectors implies that the transposes of $\boldsymbol{\Gamma}$ and $\mathbf{B}$ appear in equation (2).

The path coefficients in $\boldsymbol{\Gamma}$ and $\mathbf{B}$ are the parameters of primary interest. They describe the composites' interrelations.The weights are only necessary for model estimation. At most they give some information about the indicators' relative relevance in terms of building the composites.

From the structural model' recursiveness, it follows that $(\mathbf{I}-\mathbf{B}')$ is regular and a reduced form of the equation (2) exists:

$$\boldsymbol{\eta} = \boldsymbol{\xi}\boldsymbol{\Gamma}'(\mathbf{I}-\mathbf{B}')^{-1} + \boldsymbol{\zeta}(\mathbf{I}-\mathbf{B}')^{-1}\,. \tag{3}$$

A factor analysis point of view is often included in a composite-based structural model. In this case, the model comprising equations (5) and (2) is supplemented with a reflective part:

$$\boldsymbol{x} = \boldsymbol{\xi}\boldsymbol{\Lambda}_x' + \boldsymbol{\delta}\,, \tag{4a}$$
$$\boldsymbol{y} = \boldsymbol{\eta}\boldsymbol{\Lambda}_y' + \boldsymbol{\varepsilon}\,. \tag{4b}$$

The matrices of the loadings $\boldsymbol{\Lambda}_x$ and $\boldsymbol{\Lambda}_y$ have the same structure as $\mathbf{W}_1$ and $\mathbf{W}_2$. Henseler et al. (2014) state, that in composite-based factor models the covariance matrices of the errors are block diagonal. This makes the difference between factor based SEM's and composite-based factor models, see figure 1. This assumption is necessary to allow the loadings to be estimated by means of multivariate regression. Multivariate regression errors are correlated by the pure method. This is done especially in the PLS context. In mode A, PLS estimates the loadings by means of multivariate regression. Therefore, the errors $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ are allowed to be blockwise correlated.

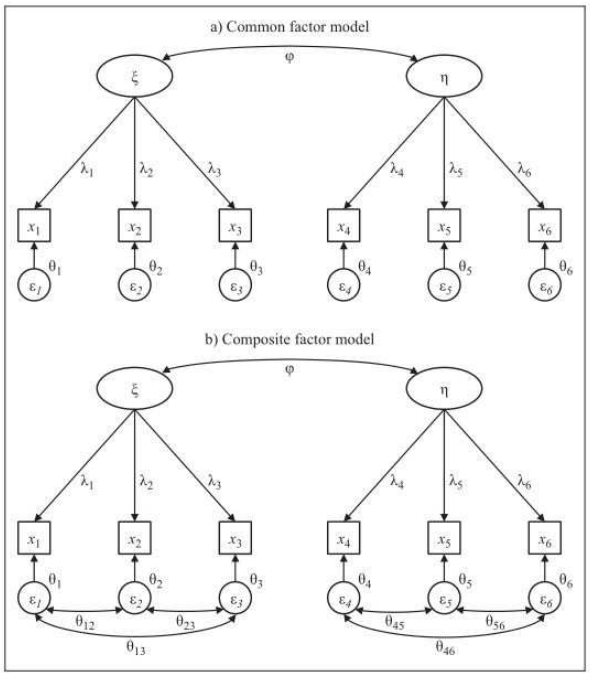

Figure 1: Contrasting common factor with composite-based factor model (Henseler et al. 2014)

A combination of a formative and a reflective model is also considered. These models are called MIMIC models in structural equation modelling's standard terminology. Only the

indicators of endogenous composites have a reflective relation. That means only equation (4b) is present but not (4a).

Models belonging to one of these three scenarios are considered only in this text. They are referenced to as standard scenarios in the following:
• Scenario ff is the formative-formative scenario. Here, no loading is included in the model.
• Scenario rr is the reflective-reflective scenario. It includes all loadings, namely the ones for the indicators of the exogenous and for the endogenous composites.
• Scenario fr is the formative-reflective scenario. Here, loadings are included only for the indicators of the endogenous composites.

The reflective-reflective scenario is given formally by the following set of equations:

$$\boldsymbol{\xi} = \boldsymbol{x}\mathbf{W}_1 \tag{5a}$$
$$\boldsymbol{\eta} = \boldsymbol{y}\mathbf{W}_2 \tag{5b}$$
$$\boldsymbol{\eta} = \boldsymbol{\xi}\boldsymbol{\Gamma}' + \boldsymbol{\eta}\mathbf{B}' + \boldsymbol{\zeta} \tag{5c}$$
$$\boldsymbol{x} = \boldsymbol{\xi}\boldsymbol{\Lambda}_x' + \boldsymbol{\delta} \tag{5d}$$
$$\boldsymbol{y} = \boldsymbol{\eta}\boldsymbol{\Lambda}_y' + \boldsymbol{\varepsilon}\,, \tag{5e}$$

In the formative-reflective scenario subequation (5d) is not present and in the formative-formative both subequations (5d) and (5e) are omitted.

1.2 The covariance matrices of GSC models

We take a constructive point of view when deriving the covariance matrices of the models. We presume that the main parameters to be controlled are 1) the path coefficients; 2) the exogenous composites' correlations; 3) the coefficients of determination for the structural regression relations; and 4) the loadings, if present.

1.2.1 The covariance matrix of the composites

The path coefficients and the coefficients of determination are related. When path coefficients are of primary concern, the coefficients of determination result from the structural model requiring uncorrelated errors. The covariance matrix of the endogenous composites, $\boldsymbol{\Sigma}_{\eta\eta}$, can be determined directly:

$$\boldsymbol{\Sigma}_{\eta\eta} = (\mathbf{I}-\mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\xi\xi}\boldsymbol{\Gamma}'(\mathbf{I}-\mathbf{B}')^{-1} + (\mathbf{I}-\mathbf{B})^{-1}\boldsymbol{\Sigma}_{\zeta\zeta}(\mathbf{I}-\mathbf{B}')^{-1} \tag{6}$$

Here, $\boldsymbol{\Sigma}_{\zeta\zeta}$ must be computed via nonlinear optimisation.

However, when specifying the coefficients of determination a priori, the path coefficients need to be determined accordingly. The computations required to determine the path coefficients only depend on the composites.

Consider the structural regression equation for the endogenous composite η_c given in (2):

$$\eta_c = \boldsymbol{\xi}\boldsymbol{\gamma}_c + \boldsymbol{\eta}_{1:c-1}\boldsymbol{\beta}'_{c,1:c-1} + \zeta_c\,, \quad 1 \le c \le q_2,$$

Here, $\boldsymbol{\beta}_{c,1:c-1}$ is the row vector consisting of the first $c-1$ elements of row c of $\mathbf{B}$. $\boldsymbol{\eta}_{1:c-1}$ is the vector of the endogenous latent variables related to rows 1 to $c-1$ of $\mathbf{B}$. The coefficients of the composites that do not appear in the regression equation of η_c are zero.

Step 1: Let $\mathbf{B}, \mathbf{\Sigma}_{\xi\xi}, q_1, q_2$ be given.
Set a start vector of length q_2 for $\mathbf{\Sigma}_{\zeta\zeta}$.
Define a function to compute
1. the right hand side of equation (6),
2. the sum of the squared differences of its diagonal elements and 1.

Step 2: Deliver the function and the start values to a nonlinear optimization routine.

Figure 2: Nonlinear determination of the matrix $\mathbf{\Sigma}_{\zeta\zeta}$.

This, together with the covariance matrix $\mathbf{\Sigma}_{(q_1+c-1),(q_1+d-1)}$ of $(\boldsymbol{\xi}, \boldsymbol{\eta}_{1:c-1})$ and $(\boldsymbol{\xi}, \boldsymbol{\eta}_{1:d-1})$, results in:

$$\mathrm{Var}(\eta_c) = (\boldsymbol{\gamma}_c, \boldsymbol{\beta}_{c,1:c-1})\mathbf{\Sigma}_{(q_1+c-1),(q_1+c-1)}(\boldsymbol{\gamma}_c, \boldsymbol{\beta}_{c,1:c-1})' + \sigma^2_{\zeta_c}, \tag{7a}$$

$$\mathrm{Cov}(\eta_c, \boldsymbol{\xi}) = (\boldsymbol{\gamma}_c, \boldsymbol{\beta}_{c,1:q_1+c-1})\mathbf{\Sigma}_{(q_1+c-1),q1}, \tag{7b}$$

$$\mathrm{Cov}(\eta_c, \eta_d) = (\boldsymbol{\gamma}_c, \boldsymbol{\beta}_{c,1:c-1})\mathbf{\Sigma}_{(q_1+c-1),(q_1+d-1)}, \quad 1 \le d \le c. \tag{7c}$$

These equations provide the relations required to compute the composites' covariance matrix. This is all one needs when the simulation is focussed on the the path coefficients.

One would usually choose $\mathbf{B}$ if one wants a specific vector $\boldsymbol{r}^2 = (R^2_1, \ldots, R^2_{q_2})$ of the coefficients of determination for the structural regressions. Thereafter, the coefficient of determination for the regression of η_c on $(\boldsymbol{\xi}, \boldsymbol{\eta}_{1:c-1})$, which is based on (7c), follows with the assumption $\mathrm{Var}(\eta_c) = 1$:

$$R^2_c = 1 - \sigma^2_{\zeta_c} = (\boldsymbol{\gamma}_c, \boldsymbol{\beta}_{c,1:c-1})\mathbf{\Sigma}_{(q_1+c-1),(q_1+c-1)}(\boldsymbol{\gamma}_c, \boldsymbol{\beta}_{c,1:c-1})'. \tag{8}$$

One has to work through matrix $\mathbf{B}$ from row q_1+1 to the last one to modify the path coefficient to reach the desired coefficients of determination. The first part of the covariance matrix is given by $\mathbf{\Sigma}_{\xi\xi}$. After the modification of the path coefficients in row q_1+c of $\mathbf{B}$, the covariance matrix of the composites must be augmented by row and column c before the coefficients of row $c+1$ can be modified.

Initially, choose the row vector $\boldsymbol{\beta}_{q_1+c}$ as preferred. Subsequently, this preliminary value is multiplied by a factor τ which makes (8) hold true:

$$\tau = \sqrt{\frac{R^2_c}{((\boldsymbol{\gamma}_c, \boldsymbol{\beta}_{c,1:c-1})\mathbf{\Sigma}_{(q_1+c-1),(q_1+c-1)}(\boldsymbol{\gamma}_c, \boldsymbol{\beta}_{c,1:c-1})'}}. \tag{9}$$

Example 1.1
We illustrate the determination of the covariance matrix and the path coefficients in the case of a given vector $\boldsymbol{r}^2$. For this purpose we consider the structural model

$$(\eta_1, \eta_2, \eta_3) = (\xi_1, \xi_2, \xi_3)\begin{pmatrix} \gamma_{11} & 0 & 0 \\ \gamma_{12} & \gamma_{22} & 0 \\ 0 & \gamma_{23} & 0 \end{pmatrix} + (\eta_1, \eta_2, \eta_3)\begin{pmatrix} 0 & 0 & \beta_{31} \\ 0 & 0 & \beta_{32} \\ 0 & 0 & 0 \end{pmatrix} + (\zeta_1, \zeta_2, \zeta_3).$$

The covariance matrix of the exogenous composites and the coefficients of determination of the regressions for the endogenous composites are set to:

$$\mathbf{\Sigma}_{\xi\xi} = \begin{pmatrix} 1 & 0.4 & 0.1 \\ 0.4 & 1 & 0.3 \\ 0.1 & 0.3 & 1 \end{pmatrix}, \quad \boldsymbol{r}^2 = \begin{pmatrix} 0.8 & 0.7 & 0.6 \end{pmatrix}.$$

The preliminary choice of the path coefficients is $\gamma_{11} = \gamma_{22} = 0.6$, $\gamma_{12} = \gamma_{23} = 0.5$, $\beta_{31} = \beta_{32} = 0.4$.

First, the regression model $\eta_1 = \gamma_{11}\xi_1 + \gamma_{12}\xi_2 + \zeta_1$ is considered. Using $\mathrm{Var}(\eta_1) = \gamma_{11}^2 + \gamma_{12}^2 + 2\gamma_{11}\gamma_{12}\mathrm{Cov}(\xi_1, \xi_2) + \mathrm{Var}(\zeta_1) = 1$ one obtains $\mathrm{Var}(\zeta_1) = 0.15$. In order to achieve $R_1^2 = 1 - \mathrm{Var}(\zeta_1) = 0.8$ the coefficients γ_{11}, γ_{12} are multiplied by $\tau = \sqrt{0.8/0.85}$. The second regression model $\eta_2 = \gamma_{22}\xi_2 + \gamma_{23}\xi_3 + \zeta_2$ results in $\mathrm{Var}(\zeta_2) = 0.21$. The resulting factor is $\tau = \sqrt{0.7/0.79}$. Up to this point the modified path coefficients are: $\gamma_{11} = 0.582$, $\gamma_{12} = 0.485$, $\gamma_{22} = 0.565$, $\gamma_{23} = 0.471$.

The covariance matrix of (ξ, η_1, η_2) must be determined to compute the factor for the third regression. Formulas (7) result in:

$$\mathrm{Cov}(\eta_1, \xi) = \begin{pmatrix} 0.776 & 0.718 & 0.204 \end{pmatrix}$$
$$\mathrm{Cov}(\eta_2, \xi) = \begin{pmatrix} 0.273 & 0.706 & 0.640 \end{pmatrix}$$
$$\mathrm{Cov}(\eta_1, \eta_2) = \begin{pmatrix} 0.565 & 0.471 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0.4 & 0.1 & 0.776 \\ 0.4 & 1 & 0.3 & 0.718 \\ 0.1 & 0.3 & 1 & 0.204 \end{pmatrix} \begin{pmatrix} 0 \\ 0.565 \\ 0.471 \\ 0 \end{pmatrix} = 0.501\,.$$

With this given covariance, one should proceed as with the first two regressions. This gives the factor $\tau = \sqrt{0.6/0.346}$. Subsequently the matrices $\mathbf{\Gamma}$ and $\mathbf{B}$ are:

$$\mathbf{\Gamma} = \begin{pmatrix} 0.582 & 0.485 & 0 \\ 0 & 0.565 & 0.471 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0.447 & 0.447 & 0 \end{pmatrix}.$$

Finally, the complete covariance matrix of the composites is computed, again using formulas (7):

$$\begin{pmatrix} 1.000 & 0.4 & 0.1 & 0.776 & 0.273 & 0.469 \\ 0.4 & 1.000 & 0.3 & 0.718 & 0.706 & 0.637 \\ 0.1 & 0.3 & 1.000 & 0.204 & 0.640 & 0.377 \\ 0.776 & 0.718 & 0.204 & 1.000 & 0.501 & 0.671 \\ 0.273 & 0.706 & 0.640 & 0.501 & 1.000 & 0.671 \\ 0.469 & 0.637 & 0.377 & 0.671 & 0.671 & 1.000 \end{pmatrix}$$

1.2.2 The covariance matrix of the indicators in scenario ff

Scenario ff is the formative model. From the model equations we derive

$$\mathbf{\Sigma}_{\xi\xi} = \mathbf{W}_1' \mathbf{\Sigma}_{xx} \mathbf{W}_1 \tag{10a}$$
$$\mathbf{\Sigma}_{\eta\eta} = \mathbf{W}_2' \mathbf{\Sigma}_{yy} \mathbf{W}_2 \tag{10b}$$
$$\mathbf{\Sigma}_{\eta\eta} = (\mathbf{I} - \mathbf{B})^{-1} \mathbf{\Gamma} \mathbf{\Sigma}_{\xi\xi} \mathbf{\Gamma}' (\mathbf{I} - \mathbf{B}')^{-1} + (\mathbf{I} - \mathbf{B})^{-1} \mathbf{\Sigma}_{\zeta\zeta} (\mathbf{I} - \mathbf{B}')^{-1} \tag{10c}$$

With a choice of $\mathbf{\Sigma}_{\xi\xi}$, the covariance matrix of the $\boldsymbol{x}$-indicators and the weights $\mathbf{W}_1$ must be determined so that (10a) is fulfilled.

One has a great degree of freedom to choose $\mathbf{\Sigma}_{\boldsymbol{xx}}$ and the standardised weights, resulting in a given $\mathbf{\Sigma}_{\xi\xi}$. First, each block of indicators of the different exogenous composites can be dealt with separately, with only the standardisation of the composites needing to be ensured. This means that $\xi_g = \boldsymbol{x}_g\mathbf{w}_g$, $\mathbf{w}_g'\mathbf{\Sigma}_{\boldsymbol{x}_g\boldsymbol{x}_g}\mathbf{w}_g = 1$ must be fulfilled. This can, for example, be achieved by setting $\mathbf{\Sigma}_{\boldsymbol{x}_g\boldsymbol{x}_g}$ as the identity matrix and choosing the weights vectors such that $\mathbf{w}_g'\mathbf{w}_g = 1$. However, this covariance matrix can be chosen arbitrarily and subsequently scaled to fulfil equation (15a). If the exogenous composites are assumed to be uncorrelated, one uses $\mathbf{\Sigma}_{\boldsymbol{x}_g\boldsymbol{x}_h} = \mathbf{0}$ for $g \neq h$. If two composites are correlated, the correlations between the different blocks' indicators must be set appropriately. An easy way of doing this is to preset $\mathbf{\Sigma}_{\boldsymbol{x}_g\boldsymbol{x}_h}$ and to scale it such that $\mathbf{w}_g'\mathbf{\Sigma}_{\boldsymbol{x}_g\boldsymbol{x}_h}\mathbf{w}_h = \sigma_{\xi_g\xi_h}$. Becker, Rai, and Rigdon (2013) first used this approach in a specific situation.

In the next step, $\mathbf{B}$ is given, or must be determined according to the given vector $\boldsymbol{r}^2$ of the coefficients of determination (see section 1.2.1). Thereafter it is possible to obtain $\mathbf{\Sigma}_{\eta\eta}$ as described in section 1.2.1. $\mathbf{\Sigma}_{\boldsymbol{yy}}$ and the weights $\mathbf{W}_2$ are determined in the same way as the covariance matrix of the X-indicators, using (10b).

The covariances of the exogenous and the endogenous composites can be used to determine $\mathbf{\Sigma}_{\boldsymbol{xy}}$. First, from (5) it follows:

$$\mathbf{\Sigma}_{\xi\eta} = \mathbf{W}_1'\mathbf{\Sigma}_{\boldsymbol{xy}}\mathbf{W}_2 \tag{11}$$

whereas (3) leads to:

$$\mathbf{\Sigma}_{\xi\eta} = \mathbf{\Sigma}_{\xi\xi}\mathbf{\Gamma}'(\mathbf{I}-\mathbf{B}')^{-1}. \tag{12}$$

The combination of these two equations provides a necessary condition that must be fulfilled:

$$\mathbf{W}_1'\mathbf{\Sigma}_{\boldsymbol{xy}}\mathbf{W}_2 = \mathbf{\Sigma}_{\xi\xi}\mathbf{\Gamma}'(\mathbf{I}-\mathbf{B}')^{-1}. \tag{13}$$

Choosing the covariance matrix $\mathbf{\Sigma}_{\boldsymbol{xy}}$ as

$$\mathbf{\Sigma}_{\boldsymbol{xy}} = \mathbf{\Sigma}_{\boldsymbol{xx}}\mathbf{W}_1\mathbf{\Gamma}'(\mathbf{I}-\mathbf{B}')^{-1}\mathbf{\Sigma}_{\eta\eta}^{-1}\mathbf{W}_2'\mathbf{\Sigma}_{\boldsymbol{yy}} \tag{14}$$

makes (13) hold true. To reach this result, one has to insert this expression into the left-hand side of (13) and to consider the relations for the covariance matrices of the composites.

A quasi-code for the computation of the covariance matrices of the indicators is given in figure 3.

Step 1: Choose $\mathbf{\Sigma}_{\xi\xi}$, $\mathbf{B}$, $\boldsymbol{r}^2 = (R_1^2, \ldots, R_{q_2}^2)$ such that for row j of $\mathbf{B}$
$R_j^2 = \mathbf{b}_j\mathrm{Var}((\boldsymbol{\xi}, \boldsymbol{\eta}))\mathbf{b}_j'$

Step 2: Choose $\mathbf{W}_1$ and $\mathbf{\Sigma}_{\boldsymbol{xx}}$ such that $\mathbf{\Sigma}_{\xi\xi} = \mathbf{W}_1'\mathbf{\Sigma}_{\boldsymbol{xx}}\mathbf{W}_1$

Step 3: Use the method of section 1.2.1 or (6) to determine $\mathbf{\Sigma}_{\eta\eta}$

Step 4: Choose $\mathbf{\Sigma}_{\boldsymbol{yy}}$ and $\mathbf{W}_2$ such that $\mathbf{\Sigma}_{\eta\eta} = \mathbf{W}_2'\mathbf{\Sigma}_{\boldsymbol{yy}}\mathbf{W}_2$

Step 5: $\mathbf{\Sigma}_{\boldsymbol{xy}} = \mathbf{\Sigma}_{\boldsymbol{xx}}\mathbf{W}_1\mathbf{\Gamma}'(\mathbf{I}-\mathbf{B}')^{-1}\mathbf{\Sigma}_{\eta\eta}^{-1}\mathbf{W}_2'\mathbf{\Sigma}_{\boldsymbol{yy}}$

Figure 3: Determination of the covariance matrices of the indicators for formative models

Example 1.2
Continuing the example 1.1 shows how to determine the covariance matrix of the indicators. With the results already obtained, step 2 of figure 3 should be taken next. Let

$$\mathbf{K}=\begin{pmatrix} 1 & 0.3 & 0.2 \\ 0.3 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{xx}}=\begin{pmatrix} \mathbf{K} & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{K} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{K} \end{pmatrix} \text{ and } \mathbf{W}_1=\begin{pmatrix} \mathbf{w}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{w}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{w}_3 \end{pmatrix}$$

where $\mathbf{1}$ is a 3×3 matrix of ones, $\mathbf{w}_1 = (0.4, 0.5, 0.6)'$ and $\mathbf{0}$ a vector of zeros. $\mathbf{w}_2$ and $\mathbf{w}_3$ are chosen suitably.

First, $\mathbf{W}_1$ has to be standardised. This is done by computing $\mathbf{w}_1/\sqrt{f}$ with $f = \mathbf{w}_1'\mathbf{K}\mathbf{w}_1 = 1.106$, and by substituting the new vector for the old $\mathbf{w}_1$. $\boldsymbol{w}_2$ and $\mathbf{w}_3$ are standardised analogously. Subsequently, blocks of ones in $\boldsymbol{\Sigma}_{\boldsymbol{xx}}$ have to be changed such that the covariances in $\boldsymbol{\Sigma}_{\xi\xi}$ are recovered. For example, to obtain $\sigma_{13} = 0.469$, the ones in the first three rows and the last three columns are modified to $0.469/(\mathbf{w}_1'\mathbf{1}\mathbf{w}_3)$.

The matrix $\mathbf{W}_2$ is dealt with analogously by using $\boldsymbol{\Sigma}_{\eta\eta}$. Finally, $\boldsymbol{\Sigma}_{\boldsymbol{xy}}$ is computed using equation (13) and the complete covariance matrix is built.

1.2.3 The covariance matrix of the indicators in scenario rr

Scenario rr is the one with composite-based factor models. The equations (5) and (3) lead to:

$$\boldsymbol{\Sigma}_{\xi\xi} = \mathbf{W}_1'\boldsymbol{\Sigma}_{\boldsymbol{xx}}\mathbf{W}_1 \tag{15a}$$
$$\boldsymbol{\Sigma}_{\eta\eta} = \mathbf{W}_2'\boldsymbol{\Sigma}_{\boldsymbol{yy}}\mathbf{W}_2 \tag{15b}$$
$$\boldsymbol{\Sigma}_{\boldsymbol{xx}} = \boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{\Sigma}_{\xi\xi}\boldsymbol{\Lambda}_{\boldsymbol{x}}' + \boldsymbol{\Sigma}_{\boldsymbol{\delta\delta}} \tag{15c}$$
$$\boldsymbol{\Sigma}_{\eta\eta} = (\mathbf{I}-\mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\xi\xi}\boldsymbol{\Gamma}'(\mathbf{I}-\mathbf{B}')^{-1} + (\mathbf{I}-\mathbf{B})^{-1}\boldsymbol{\Sigma}_{\zeta\zeta}(\mathbf{I}-\mathbf{B}')^{-1} \tag{15d}$$
$$\boldsymbol{\Sigma}_{\boldsymbol{yy}} = \boldsymbol{\Lambda}_{\boldsymbol{y}}\boldsymbol{\Sigma}_{\eta\eta}\boldsymbol{\Lambda}_{\boldsymbol{y}}' + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}} \tag{15e}$$
$$\boldsymbol{\Sigma}_{\boldsymbol{xy}} = \boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{\Sigma}_{\xi\xi}\boldsymbol{\Gamma}'(\mathbf{I}-\mathbf{B}')^{-1}\boldsymbol{\Lambda}_{\boldsymbol{y}}'\,. \tag{15f}$$

The following additional equations can be deduced from (15):

$$\boldsymbol{\Sigma}_{\xi\xi} = \mathbf{W}_1'\left(\boldsymbol{\Lambda}_{\boldsymbol{x}}\boldsymbol{\Sigma}_{\xi\xi}\boldsymbol{\Lambda}_{\boldsymbol{x}}' + \boldsymbol{\Sigma}_{\boldsymbol{\delta\delta}}\right)\mathbf{W}_1\,, \tag{16a}$$
$$\boldsymbol{\Sigma}_{\eta\eta} = \mathbf{W}_2'\left(\boldsymbol{\Lambda}_{\boldsymbol{y}}\boldsymbol{\Sigma}_{\eta\eta}\boldsymbol{\Lambda}_{\boldsymbol{y}}' + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}\right)\mathbf{W}_2\,. \tag{16b}$$

The conditions (16) can not be satisfied in many cases when the errors $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ are supposed to be uncorrelated. Then the model has intrinsic inconsistencies and can not be used as a model for a real application.

We allow blockwise correlated error vectors $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ as was stated above. This may be seen as avoiding inconsistency of the models by introducing additional parameters. It would follow the spirit of John von Neumann's statement: 'With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.' (Dyson 2004).

The following highlights the problem's relevance: In seven of the 15 examples taken from literature, the uncorrelated error vectors $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ did not allow for determing weights satisfying (16). Columns three and four of Table 1 give the values of the nonlinear optimisation criterion used to determine weights such that (16a) and (16a) were satisfied. It is concluded that no weights exist if the optimisation resulted in a value not small enough, greater than 0.01, say.

Table 1: Model check for examples from literature

No.	number of β's and λ's	values of optimisation criterion uncor. error		cor. error		Source
1	2/9	0	0.0399	0	2e-12	Hwang et al. (2010)
2	2/12	0.0241	5e-32	2e-11	1e-12	Aguirre-Urreta et al. (2013)
3	2/12	0.0026	5e-32	1e-12	1e-12	Sanchez (2013)
4	2/12	0.0026	5e-32	2e-12	2e-13	Chin & Newsted (1999)
5	3/21	5e-32	0.0164	1e-12	6e-06	Bergami & Bagozzi (2000)
6	3/18	0	0.0036	2e-12	4e-11	Eberl & v. Mitschke (2006)
7	4/12	0.0045	0.010	2e-12	7e-12	Qureshi & Compeau (2009)
8	4/30	5e-32	0.0017	1e-12	1e-05	Lu et al. (2011)
9	4/20	0.0256	0.0078	4e-11	6e-12	Dijkstra & Henseler (2015)
10	4/21	0.0008	0.0006	0.0002	7e-12	Albers & Hildebrandt (2006)
11	6/28	0	0.0224	2e-09	1e-10	Chin & Newsted (1999)
12	7/42	1e-31	0.0184	3e-13	4e-06	Aguirre-U. & Rönkkö (2017)
13	9/24	0	0.7031	5e-13	2e-09	Reinartz et al. (2009)l
14	9/23	5e-32	0.0568	3e-15	9e-09	Tenenhaus (2008)
15	10/28	0.0040	0.0515	1e-11	5e-09	Hair et al. (2017)

1.2.4 The covariance matrix of the indicators in scenario fr

This scenario is a mixture of the two scenarion already investigated. The covariance matrix of the indicators is determined therefore by following first the description of scenario ff and then that of scenario rr. This gives

$$\boldsymbol{\Sigma}_{\boldsymbol{xx}} \quad \text{with} \quad \boldsymbol{\Sigma}_{\xi\xi} = \mathbf{W}_1' \boldsymbol{\Sigma}_{\boldsymbol{xx}} \mathbf{W}_1$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{yy}} = \boldsymbol{\Lambda}_y \boldsymbol{\Sigma}_{\eta\eta} \boldsymbol{\Lambda}_y' + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}} .$$

As in scenario rr, the errors $\boldsymbol{\varepsilon}$ are supposed to be blockwise correlated.

The crosscovariance matrix of $\boldsymbol{x}$ and $\boldsymbol{y}$ can be derived from the set of equations describing this scenario:

$$\boldsymbol{\Sigma}_{\boldsymbol{xy}} = \boldsymbol{\Sigma}_{\boldsymbol{xx}} \mathbf{W}_1 \boldsymbol{\Gamma}' (\mathbf{I} - \mathbf{B}')^{-1} \boldsymbol{\Lambda}_y' .$$

2 Simulation of GSC models

The covariance matrix can be used to simulate a data set. To do so, principal-components factorisation (or any factorisation, for that matter) is performed on the population correlation matrix that is to underlie the random numbers. One random number is generated for each component to generate a multivariate random vector; each random variable is defined as the sum of the products of the variable's component loadings and the random number corresponding to each of the components. The data are normally distributed if the independent random numbers originate from a normal distribution.

It is descibed in section 2.3 how to determine the correlation matrix of the indicators in scenario ff. One has to ensure that the correlation matrix is positive definite to be used for simulation.

The weights are not required in scenario rr. But the parameters used to simulate a model must obviously be chosen such that both relations (16a) and (16b) hold true. In other words, the chosen parameters can only be used, if sets of standardised weights can be found that fulfill these relations . But the determination of the correlated errors turned out to result often in non positve definite covariance matrices which can not be factorized.

The following alternative is possible due to the blockwise nature of correlation matrices. The model is simulated with diagonal covariance matrices $\mathbf{\Sigma}_{\delta\delta}$ and $\mathbf{\Sigma}_{\varepsilon\varepsilon}$. This will affect the weights but should not affect the loadings and the path coefficient of simulated model.

A quasicode for performing simulations of a composite-based factor model is given in figure 4.

Step 1: Choose $\mathbf{\Sigma}_{\xi\xi}$, $\mathbf{B}$, $\boldsymbol{r}^2 = (R_1^2, \ldots, R_{q_2}^2)$ such that for row j of $\mathbf{B}$
$R_j^2 = \mathbf{b}_j \mathrm{Var}((\boldsymbol{\xi}, \boldsymbol{\eta}))\mathbf{b}_j'$
Choose loading matrices $\mathbf{\Lambda}_{\boldsymbol{x}}$, $\mathbf{\Lambda}_{\boldsymbol{y}}$.

Step 2: Determine $\mathbf{\Sigma}_{\boldsymbol{xx}}$ and $\mathbf{\Sigma}_{\boldsymbol{\delta\delta}}$ according (15c) such that the first one has unit diagonal.

Step 3: Use the method of section 1.2.1 or equation (6) to determine $\mathbf{\Sigma}_{\eta\eta}$

Step 4: Determine $\mathbf{\Sigma}_{\boldsymbol{yy}}$ and $\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}$ according (15e) such that the first one has unit diagonal.

Step 5: Compute the crosscovariance matrix $\mathbf{\Sigma}_{\boldsymbol{xy}}$ by using (15f).

Step 6: Built the covariance matrix of the indicators from its parts.

Step 7: Compute a factorisation of the covariance matrix and multiply it from the left with a suitably sized matrix of independent random numbers.

Figure 4: Determination of the covariance matrices of the indicators for composite-based factor models

Example 2.1
An simple example given in Sanchez (2013) is considered. It has two exogenous composites, Attac and Defense and one endogenous, Success. Each has four idicators. The paper states also the values of the path coefficients and the loadings. They are used to compute the covariance matrices. The resulting covariance matrix of the indicators does not allow a solution of the critical equations. Nevertheless (15c) (15e) and (15f) are used to built the covariance matrix to be used for simulation. The resulting covariance matrix is positve definite. Random samples can be generated with it.

```
1 library(cbsem)
2 B <- matrix(c(0,0,0,0,0,0,0.76, -0.28,0),3,3,byrow=T)
3 Sxixi <- matrix(c( 1, -0.47, -0.47, 1),2,2)
4 indicatorx <- c(1,1,1,1,2,2,2,2)
```

```
indicatory <- c(1,1,1,1)
lambdax <- c(0.83,0.84,0.86,0.94,-0.89,-0.75,0.88,0.48)
lambday <- c(0.94,0.97,0.89,0.78)
out <- gscmcov(B,indicatorx,indicatory,lambdax,lambday,wx=NULL,
                            wy=NULL,Sxixi,R2=NULL)
eigen(out$S,only.values=T)
C <- chol(out$S)
data <- matrix(rnorm(50*12),50,12)%*%C
```

There are several suggestions for generating data from nonnormal distributions with preset properties. Vale and Maurelli (1983) extended the Fleishman (1978) method to generate multivariate random numbers with specified intercorrelations and univariate means, variances, skewness values, and kurtoses. First, they produce a suitably sized matrix of independent, normally distributed random numbers. They subsequently compute the Fleishman's transformation coefficients and by using them an intermediate correlation matrix from the desired indicators' correlation matrix. A principal-components factorisation is performed of this intermediate correlation matrix and the resulting factor is multiplied with the matrix of independent normally distributed random numbers. Finally, the Fleishman transformation is applied componentwise.

Example 2.2
This method was used for a small simulation experiment to compare the estimation via the GSCA approach with PLS for formative models. Different levels of skewness $\sqrt{\beta_1}$ and excess kurtosis β_2 were chosen. The levels correspond to normal, Laplace, exponential and t_5-distributions (although the empirical values of the kurtosis are smaller than those of the target ones). Fifty samples of size $n = 100$ were generated for each distribution from the model used as the example above while the function `gscals`, see Schlittgen (2018), and the implementation `plspath` of the PLS procedure were used to estimate the model. Figure 5 shows the differences between the estimates and the path coefficients used for simulation. The estimatess are not better than in the normalsimulation methods, probably due to the multiplicative manner in which the model parameters appear in the estimation equations. Overall, the differences between the two estimation methods's results are small. However, the results by `gscals` are a bit closer to by the level even though the PLS estimates are rather more closely grouped around the true values. The shape of the distribution seems to be of no importance.

The covariance matrix **S** was determined in the two examples 1.1 and 1.2. Then the code for generating a sample of size $n = 100$ is as follows.

```
S <- rbind(cbind(Sxx,Sxy), cbind(t(Sxy),Syy ))
skew <- 0
kurt <- 6
startv <- FleishmanIC(skew, kurt)
out <- NewtonFl(c(skew,kurt),startv)
Fcoef <- out$coefficients
dat <- rValeMaurelli(100, S, Fcoef)
```

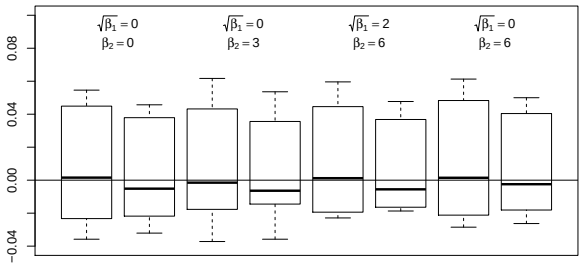

Figure 5: Deviation of estimated coefficients from model coefficients for different distributions (left: gscals, right: pls)

3 Estimation of GSC models

3.1 Reformulation of the models

The models are cast into the form introduced by Hwang and Takane (2004) to derive a method for estimation. For this, the point of view is changed to the the observations. Let $\mathbf{X}$ and $\mathbf{Y}$ be the data matrices. $\mathbf{Z}$, $\boldsymbol{\Delta}$, $\mathbf{E}$ are the matrices of scores of the error vectors $\boldsymbol{\zeta}$, $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$. Then structural and measurement parts of the model are combined into one equation.

The structural model can be written as:

$$[\mathbf{X}|\mathbf{Y}]\begin{bmatrix} \mathbf{0} \\ \mathbf{W}_2 \end{bmatrix} = [\mathbf{X}|\mathbf{Y}]\left[\begin{array}{c|c} \mathbf{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 \end{array}\right]\begin{bmatrix} \boldsymbol{\Gamma}' \\ \mathbf{B}' \end{bmatrix} + \mathbf{Z}. \tag{17}$$

The measurement model of the reflective-reflective scenario is:

$$[\mathbf{X}|\mathbf{Y}] = [\mathbf{X}|\mathbf{Y}]\left[\begin{array}{c|c} \mathbf{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 \end{array}\right]\begin{bmatrix} \boldsymbol{\Lambda}_x' & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_y' \end{bmatrix} + [\boldsymbol{\Delta}|\mathbf{E}]. \tag{18}$$

Combining these two equations lead to:

$$[\mathbf{X}|\mathbf{Y}]\left[\begin{array}{cc} \mathbf{0} & |\mathbf{I} \\ \mathbf{W}_2 & \end{array}\right] = [\mathbf{X}|\mathbf{Y}]\left[\begin{array}{c|c} \mathbf{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 \end{array}\right]\left[\begin{array}{c|cc} \boldsymbol{\Gamma}' & \boldsymbol{\Lambda}_x' & \mathbf{0} \\ \mathbf{B}' & \mathbf{0} & \boldsymbol{\Lambda}_y' \end{array}\right] + [\mathbf{Z}|\boldsymbol{\Delta}|\mathbf{E}]. \tag{19}$$

Given the weighting matrices $\mathbf{W}_1$ and $\mathbf{W}_2$, (19) states a multivariate regression relationship with the parameter matrix $\mathbf{A}$,

$$\mathbf{A} = \left[\begin{array}{c|cc} \boldsymbol{\Gamma}' & \boldsymbol{\Lambda}_x' & \mathbf{0} \\ \mathbf{B}' & \mathbf{0} & \boldsymbol{\Lambda}_y' \end{array}\right].$$

Second, the formative-reflective scenario follows, for which the measurement equation becomes:

$$[\mathbf{X}|\mathbf{Y}]\begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} = [\mathbf{X}|\mathbf{Y}]\begin{bmatrix} \mathbf{0} \\ \mathbf{W}_2 \end{bmatrix}\boldsymbol{\Lambda}_y' + \mathbf{E}. \tag{20}$$

A combination of the structural equation (17) with this measurement equation, leads to:

$$[\mathbf{X}|\mathbf{Y}]\left[\begin{array}{c|c} \mathbf{0} & \mathbf{0} \\ \mathbf{W}_2 & \mathbf{I} \end{array}\right] = [\mathbf{X}|\mathbf{Y}]\left[\begin{array}{c|c} \mathbf{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 \end{array}\right]\left[\begin{array}{c|c} \boldsymbol{\Gamma}' & \mathbf{0} \\ \mathbf{B}' & \boldsymbol{\Lambda}_y' \end{array}\right] + [\mathbf{Z}|\mathbf{E}]. \tag{21}$$

The parameter matrix $\mathbf{A}$ reduces compared to that of the reflective-reflective model scenario.

PLS knows additionaly loadings for the formative relations. In the case of formative-reflective models the equation $\boldsymbol{\xi} = \boldsymbol{x}\boldsymbol{\Lambda}_x$ has to be added. Then the measurement equation becomes

$$[\mathbf{X}|\mathbf{Y}]\left[\begin{array}{c|c} \mathbf{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{array}\right] = [\mathbf{X}|\mathbf{Y}]\left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 \end{array}\right]\left[\begin{array}{cc} \boldsymbol{\Lambda}_x & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_y' \end{array}\right] + [\boldsymbol{\Delta}|\mathbf{E}].$$

This gives together with the structural equation:

$$[\mathbf{X}|\mathbf{Y}]\left[\begin{array}{c|c|c} \mathbf{0} & \mathbf{W}_1 & \mathbf{0} \\ \mathbf{W}_2 & \mathbf{0} & \mathbf{I} \end{array}\right] = [\mathbf{X}|\mathbf{Y}]\left[\begin{array}{c|c|c} \mathbf{W}_1 & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{W}_2 \end{array}\right]\left[\begin{array}{c|c|c} \boldsymbol{\Gamma}' & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_x & \mathbf{0} \\ \mathbf{B}' & \mathbf{0} & \boldsymbol{\Lambda}_y' \end{array}\right] + [\mathbf{Z}|\boldsymbol{\Delta}|\mathbf{E}]$$

A closer inspection of this equation shows that the middle part is fulfilled without error when the weighting matrix $\mathbf{W}_1$ equals the matrix $\boldsymbol{\Lambda}_x$ of loadings. Therefore it is not necessary to determine it and the model can be reduced at the outset. This serves as additional justification for the approach pesented here.

Third, the formative-formative scenario does not state a separate measurement equation. Therefore, the model is simply given by (17). Here, $\mathbf{A} = [\boldsymbol{\Gamma}|\mathbf{B}]'$.

3.2 The estimation algorithm

The least squares methods are used to estimate the parameters in the central relationships (19), (21) and (17). Using multivariate regression means that the sum of squared residuals' target criterion should be minimized. Let the matrix on the left be $\mathbf{V}$ with which the data matrix is multiplied, and let $\mathbf{U}$ be the corresponding matrix on the right. Then, the target criterion is:

$$\operatorname{trace}((\mathbf{V}-\mathbf{U}\mathbf{A})'[\mathbf{X}|\mathbf{Y}]'[\mathbf{X}|\mathbf{Y}](\mathbf{V}-\mathbf{U}\mathbf{A})) \stackrel{!}{=} \min . \tag{22}$$

The following modification, instead of (22), measures the fit:

$$Fit = 1 - \frac{\operatorname{trace}((\mathbf{V}-\mathbf{U}\mathbf{A})'[\mathbf{X}|\mathbf{Y}]'[\mathbf{X}|\mathbf{Y}](\mathbf{V}-\mathbf{U}\mathbf{A}))}{\operatorname{trace}(\mathbf{V}'[\mathbf{X}|\mathbf{Y}]'[\mathbf{X}|\mathbf{Y}]\mathbf{V})} . \tag{23}$$

This *Fit* was proposed by Hwang and Takane (2004) and its value will be used as criterion to compare the results of the algorithms.

The proposed algorithm uses an idea taken from iteratively reweighted least squares to solve the optimization problem. It is known as W estimators for regression (Hoaglin, Mosteller and Tukey 1983). There, the unknown parameters appear also on both sides of an equation. Then the parameter values on one side are held fixed and the ones on the other side are updated. This is done here analogously. The resulting algorithm is an alternating least squares (ALS) algorithm.

Let some starting values be given. Then an update of the parameters collected in the matrix $\mathbf{A}$ is performed. The weights are fixed for that purpose. Therefore it is possible to use the relationships (19), (21) and (17) directly. It is a well-known fact that the separate estimation of every column in the parameter matrix $\mathbf{A}$ by linear regression leads to multivariate regression. It is important to be aware of the structural zeros contained in the columns. Upon selection of such a column for the estimation of its parameters, these structural zeros must be eliminated together with the corresponding columns of the regressor matrix. Eventually, the resulting estimates will lead to an updated parameter matrix $\mathbf{A}$.

Each one of the three measurement model scenarios requires a different process to update the weights. For the reflective-reflective scenario equation (19) is reformulated. With $\mathbf{B}_{r1}$ and $\mathbf{B}_{r2}$ being the submatrices of $\mathbf{B}_r$ containing the first q_1 and last q_2 columns, the new equation is:

$$[\mathbf{X}|\mathbf{Y}]\left[\begin{array}{c|c} \mathbf{0} & \\ \mathbf{W}_2 & \mathbf{I} \end{array}\right] = [\mathbf{X}|\mathbf{Y}]\left[\begin{array}{c|cc} \mathbf{W}_1\mathbf{\Gamma}' & \mathbf{W}_1\mathbf{\Lambda}_x' & \mathbf{0} \\ \mathbf{W}_2\mathbf{B}' & \mathbf{0} & \mathbf{W}_2\mathbf{\Lambda}_y' \end{array}\right] + [\mathbf{Z}|\mathbf{\Delta}|\mathbf{E}]. \tag{24}$$

The estimation of the second composited matrix to the right of equation (24) is done on the basis of that equation in the same way as $\mathbf{A}$, by using multivariate linear regressions. The estimated matrix is denoted by $\mathbf{A}^*$.

(19) and (24) result in:

$$\mathbf{A}^* \approx \left[\begin{array}{c|c} \mathbf{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 \end{array}\right]\mathbf{A}.$$

Here, $\mathbf{A}$ is the already updated matrix of parameter estimates. This leads to the following equation, allowing for new $\mathbf{W}_1$ and $\mathbf{W}_2$ estimates by multivariate regression:

$$(\mathbf{A}^*)' = \mathbf{A}'\left[\begin{array}{c|c} \mathbf{W}_1' & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2' \end{array}\right] + \mathbf{F}. \tag{25}$$

A column-wise standardization of the weights determines the updating of the parameters and weights to ensure that the latent variables have unit variance.

The updating of $\mathbf{A}$, $\mathbf{W}_1$ and $\mathbf{W}_2$ stops when the changes in their values are small enough.

Example 3.1

The reflective ECSI model is considered., cf. Schlittgen (2018). The data are Tenenhaus' mobile telefon data.

```
library(cbsem)
data(mobi250)
ind <- c(1,1,1,4,4,4,2,2,2,3,3,5,5,5,6,6,6,7,1,1,4,4,4,4)
o <- order(ind)
indicatorx <- c(1,1,1,1,1)
indicatory <- c(1,1,1,2,2,3,3,3,3,3,3,4,4,4,5,5,5)
dat <- mobi250[,o]
dat <- dat[,-ncol(dat)]
B <- matrix(c(0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,
              0,1,1,0,0,0,0,1,1,1,0,0,1,0,0,0,1,0),6,6,byrow=TRUE)
out <- gscals(dat,B,indicatorx,indicatory,loadingx=TRUE,loadingy=
              TRUE,maxiter=200,biascor=FALSE)
```

The same applies to the formative-reflective scenario. First, $\mathbf{A}$ is updated by using equation (21). Subsequently, the actualization of the weights requires a reformulation of equation (21):

$$[\mathbf{X}|\mathbf{Y}]\left[\begin{array}{c|c} \mathbf{0} & \mathbf{0} \\ \mathbf{W}_2 & \mathbf{I} \end{array}\right] = [\mathbf{X}|\mathbf{Y}]\left[\begin{array}{c|c} \mathbf{W}_1\mathbf{\Gamma}' & \mathbf{0} \\ \mathbf{W}_2\mathbf{B}' & \mathbf{W}_2\mathbf{\Lambda}_y' \end{array}\right] + [\mathbf{Z}|\mathbf{E}]. \tag{26}$$

To improve the weights, the matrix

$$\left[\begin{array}{c|c} \mathbf{W}_1\mathbf{\Gamma}' & \mathbf{0} \\ \mathbf{W}_2\mathbf{B}' & \mathbf{W}_2\mathbf{\Lambda}_y' \end{array}\right]$$

is estimated by using (26) just like in the reflective-reflective scenario. $\mathbf{A}^*$ again denotes the estimated matrix. The regression equation for estimating the weights then equals (25).

As in the reflective-reflective scenario, every $\mathbf{W}_1$ and $\mathbf{W}_2$ update leads to a standardization of the weights.

Example 3.2
An example from Ringle is taken to illustrate the formative-reflective scenario fr.

```
library(cbsem)
dat <- read.table("RingleFormReflDat.txt",header=T)
B0 <-matrix(c(0,0,0,0,0,0,0,0,0,0,
                       0,0,0,0,0,1,1,1,0,0,
                       0,0,0,1,0),5,5,byrow=T)
indicatorx <- c(1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 3)
indicatory <- c(1,1,1,2,2,2)
out <- gscals(dat,B0,indicatorx,indicatory,loadingx=FALSE,loadingy
                      =TRUE,maxiter=200,biascor=FALSE)
```

In the formative-formative scenario, only the structural equation (17) is available for improving the starting values. To update the weights, first the matrix $\mathbf{A}^*$ is estimated from the reformulated equation (17):

$$[\mathbf{X}|\mathbf{Y}]\begin{bmatrix} \mathbf{0} \\ \mathbf{W}_2 \end{bmatrix} = [\mathbf{X}|\mathbf{Y}]\begin{bmatrix} \mathbf{W}_1\boldsymbol{\Gamma}' \\ \mathbf{W}_2\mathbf{B}' \end{bmatrix} + \mathbf{Z}. \tag{27}$$

Then, the $\mathbf{W}_1$ matrix of weights is updated with the help of equation (25) in the manner described above. This is not possible for $\mathbf{W}_2$, because the last rows of $\mathbf{A} = \mathbf{B}'_r$ contain zeros only. The number of these rows equals the number of η's having no latent variables depending on them in the structural equation. This results in a block of zeros in the last rows of $\mathbf{W}_2\mathbf{B}'_{r2}$ and the weights for the corresponding endogenous latent variables can not be determined with the type of regression used above.

The matrix $\mathbf{W}_2$ of weights is therefore updated differently. Reformulating equation (17) again, we have:

$$[\mathbf{X}|\mathbf{Y}]\left[\begin{array}{c|c} \mathbf{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 \end{array}\right]\begin{bmatrix} \boldsymbol{\Gamma}' \\ \mathbf{B}' \end{bmatrix} = \mathbf{Y}\mathbf{W}_2 - \mathbf{Z}. \tag{28}$$

The matrix on the left hand side is the $n \times q_2$ score matrix of the endogenous composites. By regarding it as data matrix of dependent variables and $\mathbf{Y}$ as one of explanatory ones, we obtain new values for the entries of $\mathbf{W}_2$ by performing this multivariate regression.

Example 3.3
Albers model is consider in the case of scenario ff.

```
B0 <- matrix(
     c(0,0,0,0,0,0,0,0,0,0,0,0,0,
        0,0,0,0,0,0,1,1,0,0,0,0,
        0,1,1,0,0,0,0,0,0,1,1,0),6,6,byrow=T)
indicatorx <- c(1,1,1,2,2,2,3,3,3)
indicatory <- c(1,1,1,2,2,2,3,3,3)
out1 <- gscals(dat,B0,indicatorx,indicatory,loadingx=FALSE,loadingy=
             FALSE,maxiter=200,biascor=FALSE)
```

Step 0: Set the model with 0-1 matrices $\mathbf{W}_1, \mathbf{W}_2, \mathbf{\Gamma}, \mathbf{B}, \mathbf{\Lambda}_x, \mathbf{\Lambda}_y$.
Substitute the ones with uniform random numbers.
Choose the scenario: rr, rf, or ff.
Set Δ for convergence criterion (i.e the maximal allowed absolute difference of estimated parameters between two iterations).

Do loop:

Step 1: Update the parameters in matrix $\mathbf{A}$ by least squares using one of the equations (19), (21) or (17) according to the scenarios ff, fr, rf.

Step 2: Estimate the matrix $\mathbf{A}^*$ by least squares using one of the equations (24), (26) or (27) according to the scenarios ff, fr, rf.

Step 3: Update the weights $\mathbf{W}_1$ using the regression equation (25).
Standardize the weights in matrix $\mathbf{W_1}$

Step 4: *If* scenario = rr or = fr: Update the weights $\mathbf{W_2}$ using the regression equation (25).
If scenario = ff: Update the weights $\mathbf{W}_2$ using the regression equation (28).
Standardize the weights in matrix $\mathbf{W}_2$

Step 5: Compute the differences of the updated parameters and their values of the last round.
If The maximum of absolute differences is greater than Δ.
Go to *Step 1*.
Else Output of the estimated parameters.

Stop.

Figure 6: The gscals algorithm.

Figure 6 gives a quasi-code of our ALS-algorithm. The algorithm by Hwang and Takane (2004) is an ALS too. As they state, in general ALS can be viewed as a special type of the FP algorithm where the fixed point is a stationary point of a function to be optimized. But no investigation of theoretical aspects of our algorithm has be done yet. Nevertheless, the stationary point is characterized by the following fact: The solutions of (24) ((26) / (17)) with inserted solutions of (25) ((25)/(27) and (28)) do not change anymore and the same holds when the two equations are exchanged.

Experiments showed that the results do not depend on the choice of starting values. Therefore, uniformly distributed random variates are chosen for them. When PLS-estimates were used as starting values for the four empirical examples, the number of iterations were 21, 37, 15, 14 for the PLS-starting values and 21, 44, 16, 15 for the random starts. But the PLS-iterations must be taken into account additionally.

Empirical evidence of monotonic decrease of the maximum of absolute differences of parameters during the iterations can be given only. It was violated only two times in the first step from the initial parameter settings to the first improved estimates. This was caused perhaps by a special constellation of the random start values.

Schlittgen (2018) shows that the proposed algorithm works well compared to `matrixpls` and `GeSCA`.

3.3 Bootstrap bias correction

GSCA and PLS estimates are known to give biased results. This also holds true for the recent modification of the PLS algorithm, making its estimation consistent for reflective models (model A) (Dijkstra & Henseler 2015). The modification estimates the loadings much better, but improves the estimates of the path coefficients only a little bit in smaller samples. Thereforeit is worthwhile to investigate bootstrap bias correction. The idea behind this correction can be described concisely as follows.

Let θ be the parameter to be estimated and let $\hat{\theta}$ be an estimator of it. It may be necessary to adjust it in an additive way. Then $\hat{\theta}+t$ is the bias corrected estimator. Ideally, one would like to choose t to reduce the bias to zero, i.e. to solve $\mathrm{E}(\hat{\theta}-\theta+t)=0$. With the bootstrap, an empirical version is produced that mimics this theoretical relation. The estimate $\hat{\theta}$ computed from the sample takes the role of the theoretical parameter and the average of the bootstrap estimates $\overline{\hat{\theta}^*}$ takes the role of the mean value of the estimator. Then the bias correction fulfills

$$\overline{\hat{\theta}^*}-\hat{\theta}+t=0, \qquad \text{or} \qquad t=\hat{\theta}-\overline{\hat{\theta}^*}.$$

Therefore, the (additively) bias corrected estimator is given by

$$\hat{\theta}+t=2\cdot\hat{\theta}-\overline{\hat{\theta}^*}. \tag{29}$$

Two variants of this approach have been implemented. First, the multivariate observations were resampled and the parameters are estimated with the resampled data set. The second variant is a parametric bootstrap. With the estimated parameters, the indicators' covariance matrix is determined. Normally distributed samples were simulated with it of the same size as the original sample. These samples were used again to get botstap estimates.

A simulation study was performed to investigate the benefit of it. First, the reflective-reflective Bergami-Bagozzi model (Bergami & Bagozzi 2000), see figure 7 was considered. We considered three levels of sample size ($n=25,100,400$), and generated 500 samples for each. Three distributions were taken into account: the normal distribution, a leptocitic and a skewed distribution.

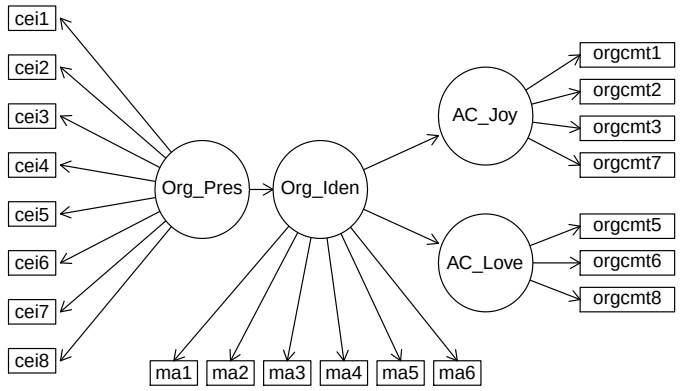

Figure 7: The specified structural equation model for Bergami and Bagozzi's organizational identification data

With the parameters estimated by the gscals algorithm the model was simulated 500 times and estimated. Estimation was done without and with bootstrap-bias correction. To give an overall impressoÃ⬡on of the relative benefit of bootstrap bias correction, the means of the absolute errors over all path coefficients and over all loadings are presented in table 2.

Table 2: Mean values of bias and mean squared error for the estimates for the Bergami-Bagozzi model

			mean values of			
distribution	n	bias cor.	bias(β)	bias(λ)	MSE(β)	MSE(λ)
normal	25	no	0.0334	0.0346	0.0327	0.0081
skew = 0,		resamp.	0.0346	0.0416	0.0333	0.0084
kurt = 0		param.	0.0062	0.0099	0.0366	0.0099
	100	no	0.0376	0.0382	0.0082	0.0033
		resamp.	0.0448	0.0399	0.0092	0.0034
		param.	0.0067	0.0105	0.0077	0.0024
	400	resamp.	0.0401	0.0391	0.0035	0.0023
		param.	0.0443	0.0386	0.0038	0.0023
		yes	0.0035	0.0101	0.0020	0.0007
skew = 0,	25	no	0.0320	0.0347	0.0315	0.0086
kurt = 3		param.	0.0025	0.0102	0.0357	0.0105
	100	no	0.0384	0.0387	0.0090	0.0035
		param.	0.0033	0.0100	0.0087	0.0027
	400	no	0.0444	0.0389	0.0038	0.0024
		param.	0.0048	0.0101	0.0020	0.0008
skew = 2,	25	no	0.0184	0.0332	0.0322	0.0119
kurt = 6		param.	0.0132	0.0087	0.0367	0.0153
	100	no	0.0394	0.0376	0.0098	0.0043
		param.	0.0020	0.0117	0.0094	0.0041
	400	no	0.0413	0.0384	0.0039	0.0026
		param.	0.0032	0.0102	0.0024	0.0012

The figures of the table show that the resampling variant of bootstrap bias correction does not work. But the parametric variant has a remarkable effect. Because of that only this is considered in the following.

The distribution does not affect neither the level of the bias nor the amount of the correction. The same holds for the sample size. But the mean squared errors indicate that the gain by reducing the bias is greater for larger n. This also holds true for normally and non-normally distributed data, despite the bootstrap also having been performed with normally distributed bootstrap samples in the latter situations. Altogether bootstrap bias correction works well for all distributions.

Now, a formative-formative model will be investigated. We use the ECSI-model with Tenenhaus' mobile phone data. A discussion in Gudergan, Ringle, Wende & Will (2008) inspired this. These authors draw attention to the possibility that the usual reflective-reflective measurement relationships are misspecifications.

First, the model is estimated. The resulting parameters $\hat{\boldsymbol{\Sigma}}_{\xi\xi}$, $\hat{\mathbf{B}}$, $\hat{\mathbf{w}}_x$, and $\hat{\mathbf{w}}_y$ are used to

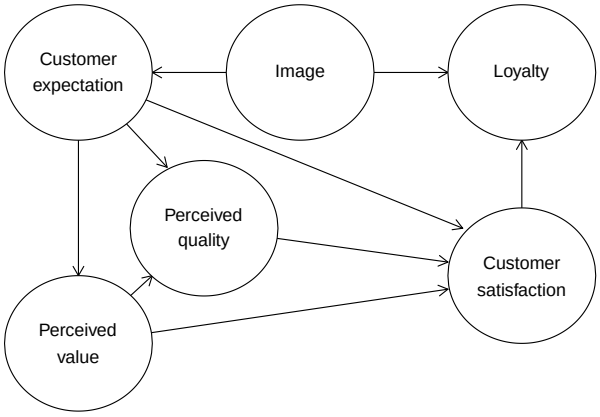

Figure 8: Stuctural model for customer satisfaction

simulate the model.

Table 3: Mean values of bias and mean squared error for the estimates for the formative ECSI model

n	bias cor.	bias(β)	MSE(β)
25	no	0.1379	0.1415
	yes	0.1224	0.2051
100	no	0.0404	0.0044
	yes	0.0340	0.0050
400	no	0.0386	0.0047
	yes	0.0390	0.0043

The 500 replications for $n = 25$ produced only 34 where the covariance matrix derived from the estimate could be used to generate bootsrap samples. With $n = 400$ this number increased to 491.

Additionally, the formative Albers-model was investigated.

Table 4: Mean values of bias and mean squared error for the estimates for the formative Albers model

n	bias cor.	bias(β)	MSE(β)
25	no	0.0595	0.0635
	yes	0.0612	0.0860
100	no	0.0150	0.0056
	yes	0.0086	0.0055
400	no	0.0115	0.0014
	yes	0.0097	0.0015

Again, a certain sample size is necessary until bootstrap bias correction has a positive effect. It does not work well for samples too small. Altogether, the bias correction works

well but not in that amount as in the reflective-reflective situation.

We use simulated data to explore the usefulness of bootstap bias correction when the model is formative-reflective. Figure 9 shows the model taken from the literature (Ringle & al. 2009). The parameter values are given in Figure 9.

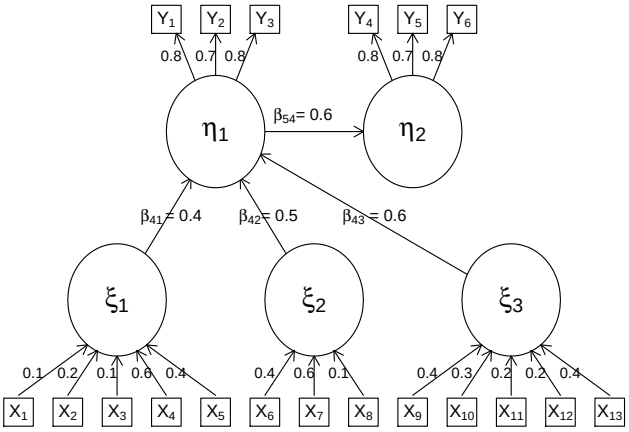

Figure 9: Ringle's model for investigating a formative-reflective scenario

The results for normally distributed data is presented in Table 5. With regard to the path coefficients the bias correction does not show a clear improvement, but ambiguous results. But there is a tendency that it pays for larger sample sizes. The bias correction has contra positive effect for small samples ($n = 25$) but positive one for larger samples.

Table 5: Mean values of bias and mean squared error for the estimates for the formative-reflective Ringle model

n	bias cor.	bias(β)	bias(λ)	MSE(β)	MSE(λ)
25	no	0.0234	0.0109	0.0903	0.0003
	yes	0.0688	0.0078	0.2386	0.0004
100	no	0.0159	0.0128	0.0151	0.0002
	yes	0.0127	0.0052	0.0607	0.0001
400	no	0.0109	0.0134	0.0023	0.0002
	yes	0.0042	0.0046	0.0017	0.0000

4 Segmentation of GSC models

4.1 An algorithm for known number of segments

Uncovering unobserved heterogeneity is a requirement to obtain valid results when using the structural equation modeling method with empirical data. Conventional segmentation methods usually fail in SEM since they account for the observations but not the composites and their relationships in the structural model.

Finding the best segmentation solution for a goal criterion is a combinatorial data assignment problem. The complexity of the problem increases exponentially with higher num-

bers of observations and/or higher numbers of segments (Cowgill, Harvey, & Watson, 1999). Conventional segmentation methods usually fail in SEM since they account for the observations but not the latent variables and their relationships in the structural model. The GSC-IRRS approach builds on an idea introduced by Schlittgen (2011) for clusterwise robust regression. In robust regression, M-estimators down-weight observations with extreme values of the dependent variable. Thereby, they mitigate the influence of outliers in the data set. One method to compute M-estimators is iteratively reweighted least squares. The weights are determined by the residuals and the larger the residuals, the smaller the weights. Since the parameters of GSC models are estimated basically by a system of least squares regressions, it is possible to use the idea of robust regression for determining a group of data and to address the segmentation problem. To adapt this idea for GSCA segmentation, outliers are not treated as such but as their own segment. Hence, when robust regression identifies a group of similar outliers, they may become a data group of their own and represent a segment-specific GSCA solution. On the other hand, within a group of data, a M-estimator down-weights inhomogeneous observations when returning the segment-specific GSCA solution.

We start with a random choice of weights v_{ik}, where i indicates an observation and $k = 1,, g$ the different segments ($\sum_{k=1}^{g} v_{ik} = 1$ for all $i = 1,\ldots,n$. Next, the method determines the segment-specific GSC solutions accounting for the weights $(v_{1k},\ldots,v_{nk})'$ of the g segment vectors in the `gscals` algorithm, by multiplying the data matrix with the square root of the weights.

Building on these results, in the next step, GSC-IRRS computes new weights. They are based on the structural model residuals (i.e., r_{ikj}) which are obtained from the g models when applied to the unweighted observations. j stands for the different regression equations in the structural model. More precisely, let $r_{ik}^2 = \sum_j r_{ikj}^2$. Then the normed reciprocal values $1/r_{ik}^2$ are used as new weights v_{ik}. Therefore an observation i gets a higher (lower) weight in segments where the sum of its squared residuals is small (large). Using these new weights as input, GSC-IRRS updates the segment-specific solutions and, again, determines new weights.

The algorithm terminates when the parameter estimates stabilize (i.e., difference of estimated coefficients between two iterations reaches a value that is smaller than a pre-defined level Δ).

Example 4.1
Scholing & Timmermann (2000) studied the interdependence in the development of certain specific economic liberties on the one hand and political rights on the other. The question was whether economic liberties can exist independently of political rights, i.e. whether private property, freedom to run a business or to choose a job, contractual freedom and freedom of pricing can be utilized merely as regulation and discovery mechanisms. In other words: whether an authoritarian political system can continue to exist side by side with economic liberties. They had data for 91 states in 1975 and 1995 on the following variables:

Competition of Parties	X_1	CP75	Y_1	CP95
Politial Rights	X_2	PR75	Y_2	PR95
Civil Liberties	X_3	CL75	Y_3	CL95
Amount of Privatisation	X_4	AoP75	Y_4	AoP95
Freedom of Foreign Exchange	X_5	FFE75	Y_5	FFE95
Freedom of Capital Movements	X_6	FCM75	Y_6	FCM95

Step 0: Set n for the number of observations;
set g for the number of groups;
set Δ for the convergence criterion, i. e. maximum difference of estimated parameters between two iterations;
set *Stop* (i.e., maximum number of generated GSC-IRRS solutions).
Randomly generate weights $\nu_{ik} \geq 0$ with $\sum_k \nu_{ik} = 1$ for all $i = 1, \ldots, n$ whereby i indicates an observation and $k = 1, \ldots, g$ the different segnments.

Do loop

Step 1: For $k = 1, \ldots, g$: Estimate the GSC path model with the ν_{ik} weighted observations.

Step 2: Determine the residuals r_{ikj} of the estimated structural regressions j using the unweighted observations.

Step 3: For each $i = 1, \ldots, n$, compute the sum of the squared values $r_{ik}^2 = \sum_j r_{ikj}^2$.

Step 4: Let the normed reciprocal values $1/r_{ik}^2$ become the new weights.

Step 5: Compare the estimated coefficients with those of the previous iteration.
If the difference is larger than Δ and the number of iteration is less than *Stop*.
Go to *Step 1*
Else
Use the maximum weight ν_{ik} to assign each observation i to a segment k.

Step 6: Compute the average value of the weighted coefficients of determination to assess and compare the quality of segmentation results.

Stop loop

Step 7: Select the final segmentation solution based on the maximum value.

Figure 10: The GSC-IRRS Algorithm

They were used as indicators in a structural model. The composite based factor model has four composites. The exogenous ones are ξ_1 Political Freedom and ξ_2 Economical Freedom 1975, endogenous are the same but for 1995.

Scholing explained in a personal communication that he was not very satisfied with the fit of the model to the data. In fact, a three cluster solution led to the following covariance matrices of the manifest variables:

Cluster 1 ($n_1 = 45$)

	CP75	PR75	CL75	AoP75	FFE75	FCM75	CP95	PR95	CL95	AoP95	FFE95	FCM95
CP75	1.00	0.91	0.82	0.32	0.25	0.45	0.77	0.83	0.81	0.39	0.44	0.69
PR75	0.91	1.00	0.94	0.24	0.33	0.49	0.76	0.83	0.85	0.34	0.52	0.79
CL75	0.82	0.94	1.00	0.19	0.35	0.45	0.75	0.83	0.86	0.33	0.56	0.75
AoP75	0.32	0.24	0.19	1.00	0.40	0.49	0.09	0.18	0.14	0.79	0.02	0.36
FFE75	0.25	0.33	0.35	0.40	1.00	0.46	-0.05	0.12	0.23	0.50	0.41	0.43
FCM75	0.45	0.49	0.45	0.49	0.46	1.00	0.25	0.38	0.36	0.46	0.32	0.64
CP95	0.77	0.76	0.75	0.09	-0.05	0.25	1.00	0.81	0.76	0.22	0.23	0.53
PR95	0.83	0.83	0.83	0.18	0.12	0.38	0.81	1.00	0.91	0.33	0.40	0.59
CL95	0.81	0.85	0.86	0.14	0.23	0.36	0.76	0.91	1.00	0.33	0.54	0.64
AoP95	0.39	0.34	0.33	0.79	0.50	0.46	0.22	0.33	0.33	1.00	0.28	0.45
FFE95	0.44	0.52	0.56	0.02	0.41	0.32	0.23	0.40	0.54	0.28	1.00	0.55
FCM95	0.69	0.79	0.75	0.36	0.43	0.64	0.53	0.59	0.64	0.45	0.55	1.00

Cluster 2 ($n_2 = 30$)

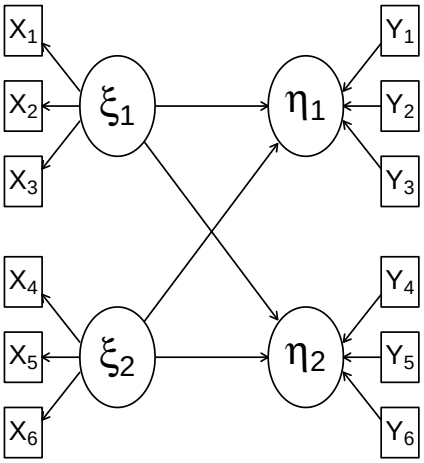

Figure 11: SEM of political and economical freedom

	CP75	PR75	CL75	AoP75	FFE75	FCM75	CP95	PR95	CL95	AoP95	FFE95	FCM95
CP75	1.00	0.93	0.79	-0.05	-0.03	-0.09	0.49	0.40	0.32	0.23	0.29	0.38
PR75	0.93	1.00	0.91	-0.05	-0.03	0.04	0.45	0.39	0.32	0.23	0.21	0.51
CL75	0.79	0.91	1.00	-0.05	0.09	0.05	0.40	0.40	0.32	0.23	0.19	0.53
AoP75	-0.05	-0.05	-0.05	1.00	0.40	0.18	0.20	0.36	0.49	0.78	0.46	0.37
FFE75	-0.03	-0.03	0.09	0.40	1.00	0.39	0.42	0.45	0.56	0.36	0.61	0.48
FCM75	-0.09	0.04	0.05	0.18	0.39	1.00	0.32	0.49	0.45	0.16	0.41	0.69
CP95	0.49	0.45	0.40	0.20	0.42	0.32	1.00	0.63	0.66	0.31	0.52	0.54
PR95	0.40	0.39	0.40	0.36	0.45	0.49	0.63	1.00	0.86	0.45	0.66	0.70
CL95	0.32	0.32	0.32	0.49	0.56	0.45	0.66	0.86	1.00	0.60	0.74	0.73
AoP95	0.23	0.23	0.23	0.78	0.36	0.16	0.31	0.45	0.60	1.00	0.64	0.45
FFE95	0.29	0.21	0.19	0.46	0.61	0.41	0.52	0.66	0.74	0.64	1.00	0.54
FCM95	0.38	0.51	0.53	0.37	0.48	0.69	0.54	0.70	0.73	0.45	0.54	1.00

Cluster 3 ($n_2 = 16$)

	CP75	PR75	CL75	AoP75	FFE75	FCM75	CP95	PR95	CL95	AoP95	FFE95	FCM95
CP75	1.00	0.85	0.71	0.55	0.61	0.60	0.22	0.05	0.06	-0.07	-0.52	-0.74
PR75	0.85	1.00	0.86	0.54	0.52	0.65	0.21	0.04	0.06	-0.12	-0.67	-0.57
CL75	0.71	0.86	1.00	0.50	0.32	0.50	0.15	0.04	0.22	0.01	-0.71	-0.54
AoP75	0.55	0.54	0.50	1.00	0.60	0.68	-0.35	-0.62	-0.49	0.54	-0.38	-0.07
FFE75	0.61	0.52	0.32	0.60	1.00	0.82	-0.06	-0.13	-0.22	0.14	-0.48	-0.16
FCM75	0.60	0.65	0.50	0.68	0.82	1.00	-0.04	-0.34	-0.35	0.01	-0.64	-0.09
CP95	0.22	0.21	0.15	-0.35	-0.06	-0.04	1.00	0.63	0.48	-0.55	-0.35	-0.68
PR95	0.05	0.04	0.04	-0.62	-0.13	-0.34	0.63	1.00	0.88	-0.53	-0.15	-0.48
CL95	0.06	0.06	0.22	-0.49	-0.22	-0.35	0.48	0.88	1.00	-0.29	-0.19	-0.51
AoP95	-0.07	-0.12	0.01	0.54	0.14	0.01	-0.55	-0.53	-0.29	1.00	0.16	0.19
FFE95	-0.52	-0.67	-0.71	-0.38	-0.48	-0.64	-0.35	-0.15	-0.19	0.16	1.00	0.33
FCM95	-0.74	-0.57	-0.54	-0.07	-0.16	-0.09	-0.68	-0.48	-0.51	0.19	0.33	1.00

Those states belong to te first cluster, where high political freedom goes together with high economical freedom and the development of both go parallel. In cluster two there is some contradiction between the two kinds of freedom in 1975 but not more in 1995. Cluster three gathers the states with contradictory development. Low political freedom in 1975 goes together with high economical freedom in 1995 and economical freedom in 1975 is different

from that in 1995. Altogether the development differs to much to built just one model for all states.

```
library(cbsem)
dat <- data(poloecfree)
dat <- dat[,-(c(1,2)]
indicatorx <-c(1,1,1,2,2,2)
indicatory <-c(1,1,1,2,2,2)
B = matrix(c(0, 0, 0, 0,
             0, 0, 0, 0,
             1, 1, 0, 0,
             1, 1, 0, 0),4,4,byrow=T)
out <- clustergscairls(dat,B,indicatorx,indicatory,loadingx=TRUE,
                                  loadingy=TRUE,3,6,1)
```

4.2 Selection of the number of segments

A method for choosing the number of segments uses Akaike's information criterion AIC. The AIC is an estimator of the relative quality of statistical models for a given set of data. Another approach to determine a suitable number of segments uses a sequence of tests. For two partitions with g and $g+1$ segments one may consider the hypotheses that the data came from model with g segments and that they came from the model with $g+1$ segments.

We do not test these two hypotheses. Instead we consider an one-dimensional parameter θ and state as null hypothesis $H_0' : \theta = 0$, such that $\theta = R^2_{W,g+1} - R^2_{W,g}$. $R^2_{W,g}$ is the weighted average of the g averages of the coefficients of determination for the regressions in the structural model. As alternative we choose $H_1' : \theta > 0$. The reason to use one-sided hypotheses comes from the fact that the adjusted determination coefficient decreases only slowly when more regressors are included. Therefore we will use more clusters only if there is a significant improvement in the fit. The test is performed with the help of a bootstrap-confidence interval for θ. We use the basic bootstrap confidence limits with an adjustment based on the double bootstrap, see Davison and Hinkley (1997, pp. 223-226). The null hypothesis $H_0 : \theta = 0$ is rejected when the lower confidence limit is greater than zero.

The hypotheses are considered to be non-nested. Cox (1961, 1962) developed a variant of likelihood ratio test for non-nested hypotheses. This has been the basis of the development of various other tests, see Davidson and MacKinnon (2004). The problem is that the test results may not be consistent. It is possible that the tests reject both, neither, or either one of the hypotheses H_1 and H_2. This goes along with the possibility that the data generating process is different from the two models under consideration. On the other hand, since the bootstrap test employed here only evaluates if a model has a significantly higher explanatory power than the other model, there is no possibility of inconsistent results.

Example 4.2
The last example will be continued. To decide about the number of clusters two tests were performed. The index of θ_g gives the number of clusters.

$\theta_1 - \theta_2$: $[-0.2499, -0.0837]$
$\theta_2 - \theta_3$: $[-0.8620, -0.3761]$

Therefore, a two cluster solution is superior to a single model but a three cluster solution is even better. A clustering with four cluster was not considered because the clusters became too small to perform the estimation.

The code of the last example is continued. Then the first comparison is performed as follows.

```
member1 <-rep(1,91)
out <- clustergscairls(dat,B,indicatorx,indicatory,loadingx=TRUE,
                                    loadingy=TRUE,2,6,1)
member2 <- out$member
boottestgscm(dat,B,indicatorx,indicatory,loadingx=TRUE,loadingy
                        =TRUE,member1,member2,0.1,inner=FALSE)
```

References

Aguirre-Urreta M, Marakas GM, Ellis ME (2013) Measurement of Composite Reliability in Research Using Partial Least Squares: Some Issues and an Alternative Approach, The DATA BASE for Advances in Information Systems 44, 11 – 43

Aguirre-Urreta M, RÃűnkkÃű M (20017) Statistical Inference with PLSc Using Bootstrap Confidence Intervals, MIS Quarterly, 43, 1 – 52

URL www.researchgate.net/publication/315690307_Statistical_Inference_with_PLSc_ Using _Bootstrap_Confidence_Intervals

Albers S, Hildebrandt L (2006) Methodische Probleme bei der Erfolgsfaktorenforschung : Messfehler, formative versus reflektive Indikatoren und die Wahl des Strukturgleichungs-Modells. Schmalenbachs Zeitschrift fÃijr betriebswirtschaftliche Forschung 58: 2 – 33

Becker J-M, Rai A, Rigdon E (2013) Predictive Validity and Formative Measurement in Structural Equation Modeling: Embracing Practical Relevance. Proceedings of the Thirty Fourth International Conference on Information Systems.

URL http://aisel.aisnet.org/icis2013/proceedings/ResearchMethods/5/

Bergami M, Bagozzi RP (2000) Self-categorization, affective commitment and group self-esteem as distinct aspects of social identity in the organization. British Journal of Social Psychology 39, 555–577

Burlander R (2008) Customer-Reationship-Management-Systeme unter Nutzung mobiler Endgeräte. Universitäsverlag, Karlsruhe

Chin WW, Newsted PR (1999): Structural Equation Modeling Analysis with Small Samples Using Partial Least Squares in: R.H. Hoyle: Statistical strategies for small sample research 1999, Sage Publications, Thousand Oaks pp. 307 – 341

Dijkstra TK, Henseler J (2015) Consistent partial least squares path modeling. MIS Quaterly 39: 297 – 316

Eberl M, von Mitschke-Collande D (2006) Die Verträglichkeit kovarianz- und varianzbasierter Schätzverfahren fÃijr Strukturgleichungsmodelle - Eine Simulationsstudie. Münchner Betriebswirtschaftliche BeitrÃd'ge, 06/2006, München

Dyson F (2004) A meeting with Enrico Fermi. Nature 427: 297

Hair JF, Hult GTM, Ringle CM, Sarstedt M, Thiele KO (2017) Mirror, mirror on the wall: a comparative evaluation of composite-based structural equation modeling methods. Journal of the Academy of Marketing Science (JAMS). doi: 10.1007/s11747-017-0517-x.

Henseler J, Dijkstra TK, Sarstedt M, Ringle CM, Diamantopoulos A, Straub DW, Ketchen Jr. DJ, Hair JF, Hult GTM, and Calantone RJ (2014) Common Beliefs and Reality About PLS: Comments on Rönkkö and Evermann (2013) Organizational Research Methods 1-28.
doi: 10.1177/1094428114526928
Hwang H (2011) GeSCA Manual. URL: www.sem-gesca.org/GeSCA_Manual.pdf
Hwang H and Takane Y (2004) Generalized structured component analysis. Psychometrika 69: 81 – 99
Hwang H, Malhotra NK, Kim Y, Tomiuk MA, Hong S (2010) A Comparative Study on Parameter Recovery of Three Approaches to Structural Equation Modeling. Journal of Marketing Research 47: 699 – 712
Jöreskog KG, Sörbom D (1989) LISREL 7-A guide to the program and applications. 2^{nd} edition. SPSS Publications, Chicago
Jöreskog KG, Wold H. (Eds.) (1982) Systems under indirect observation: Causality - structure - prediction.North Holland, Amsterdam
Lohmöller J (1989) Latent Variable Path Modelling with Partial Least Squares. Physica, Heidelberg
Lu IRR, Kwan E, Thomas DR, Cedzynski M (2011): Two new methods for estimating structural equation models: An illustration and a comparison with two established methods Intern. J. of Research in Marketing 28 (2011) 258 – 268
Qureshi I, Compeau D (2009): Assessing Between-Group Differences in Information Systems Research: A Comparison of Covariance- and Component-Based SEM, MIS Quarterly, Vol. 33, 197 – 214
Reinartz W, Haenlein M, Henseler J (2009) An empirical comparison of the efficacy of covariance-based and variance-based SEM, Intern. J. of Research in Marketing 26, 332 – 344
Ringle CM (2005) personal communication
Sanchez G (2013) PLS Path Modeling with R. URL: http://gastonsanchez.com
Schlittgen R (2018) Estimation of generalized structured component analysis models with alternating least squares. Computational Statistics. 33, 527 – 548
Schlittgen R Ringle CM Sarstedt M Becker J-M (2016) Segmentation of PLS path models by iterative reweighted regressions, Journal of Business Research, 69, 4583-4592, http://dx.doi.org/10.1016/j.jbusres.2016.04.009
Scholing E Timmermann V (2000) Der Zusammenhang zwischen politischer und Ãűkonomischer Freiheit: Eine empirische Untersuchung, Swiss Journal of Economics and Statistics, 136, 1 – 23
Schuberth F Henseler J Dijkstra T (2018) Partial least squares path modeling using ordinal categorical indicators Quality & Quantity 52, 9 -âĂŞ 35 https://doi.org/10.1007/s11135-016-0401-7
Tenenhaus M (2008) Component-based structural equation modelling. No 887, Les Cahiers de Recherche from HEC Paris.
Wold H (1982) Soft Modeling: The Basic Design and Some Extensions. In: Jöreskog KG, Wold H (eds) Systems Under Indirect Observations: Part II. Amsterdam: North-Holland: 1 – 54