A least squares approach to latent variables extraction in formative-reflective models

Marco Fattore¹, Matteo Pelagatti²,*, Giorgio Vittadini¹

Università degli Studi di Milano-Bicocca, Via Bicocca degli Arcimboldi 8, Milano, Italy

¹Department of Quantitative Methods
²Department of Statistics
Università degli Studi di Milano-Bicocca
Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy

Abstract

In this paper, we propose a new least-squares based procedure to extract exogenous and endogenous latent variables in formative-reflective structural equation models. The procedure is a valuable alternative to PLS-PM and Lisrel; it is fully consistent with the causal structure of formative-reflective schemes and extracts both the structural parameters and the factor scores, without identification or indeterminacy problems. The algorithm can be applied to virtually any kind of formative-reflective scheme, with unidimensional and even multidimensional formative blocks. To show the effectiveness of the proposal, some simulated examples are discussed. A real data application, pertaining to customer equity management, is also provided, comparing the outputs of our approach with those of PLS-PM, which may produce inconsistent results when applied to formative-reflective schemes.

Keywords: path model, formative-reflective model, least squares, reduced rank regression, PLS-PM

*Corresponding author (Tel +39-02-64485834, Fax +39-02-64485878).
Email addresses: marco.fattore@unimib.it (Marco Fattore), matteo.pelagatti@unimib.it (Matteo Pelagatti), giorgio.vittadini@unimib.it (Giorgio Vittadini)
1. Introduction

The use of formative latent constructs in data analysis is spreading. Formative models are becoming a standard tool in socio-economic research, particularly in the fields of causal modeling and multidimensional evaluation. Although at the theoretical level the debate on the role and the use of formative models is still very fired, in practice these models are involved in many applied studies (Diamantopoulos, Riefler & Roth, 2008), making it urgent to develop new and effective statistical tools to deal with them. Despite the relevance of the topic, indeed, there are still unsolved methodological problems when dealing with structural equation models comprising formative constructs. This is particularly true for formative-reflective schemes, where manifest formative variables and manifest reflective variables are causally connected via a relational structure comprising both exogenous and endogenous latent variables. Such schemes, often used in causal modeling, are usually addressed using the Lisrel algorithm, which is affected by indeterminacy problems (Vittadini, 1989), or using the PLS-PM algorithm, which cannot handle formative relationships properly.

The limitations of the Lisrel and PLS-PM algorithms have been extensively discussed in Vittadini, Minotti, Fattore & Lovaglio (2007), where an original algorithm, called RA-PM, has also been proposed as a valuable alternative. RA-PM overcomes many of the issues of Lisrel and PLS-PM, being consistent with the formative and the reflective relationships in the model and providing unique latent scores. However, RA-PM turns out to be only conditionally optimal and it does not take into account the formative side of the model adequately. For this reason, in this paper we propose a new procedure for the extraction of exogenous and endogenous latent variables in a formative-reflective scheme, which shares the same benefits of RA-PM, but satisfies a global optimum condition as it takes into account both the formative and the reflective sides of the model in a balanced way.

In a formative-reflective scheme, the exogenous latent variables play a double role. On the one hand, they should summarize their formative blocks; on the other hand, they should mediate, via the system of endogenous latent variables, the causal relationships linking the formative side to the reflective side. Realizing this, the proposed procedure extracts the exogenous latent variables balancing between these two aspects. An interesting feature of the tool is that it provides a way to check the correct specification of the model, that is to check whether the formative and the reflective sides of the model can indeed be consistently connected through the structural equation model. This feature is very important when one realizes that latent variables can be defined in a meaningful way only embedding them in larger relational structures, as discussed later in this paper. The extraction procedure has been designed for a formative-reflective scheme with unidimensional blocks of manifest variables, but it can be easily extended and adapted to the case where blocks are multidimensional as we show in Section 3.3. The algorithm can be effortlessly implemented in any programming language with matrix algebra and numerical optimisation capabilities. We implemented it in a user-friendly Ox object class (Doornik, 2007), freely available from the corresponding author.

The paper is organized as follows. Section 2 provides some insights into the present debate on formative latent variables, focusing on the role of relational structures in the definition of latent variables; Section 3 formalises the class of formative-reflective models, introduces the optimization criterion and develops the extraction procedure. Section 4 contains some artificial examples that illustrate
how our procedure can also be used as a diagnostic tool for assessing the consistency of the model with the data. Section 5 discusses an application to real data, providing also a comparison between our procedure and the widespread technique of partial least squares path modelling (PLS-PM). Section 6 concludes.

2. Formative models: an open issue

The role and the meaning of formative latent variables in structural equation models have been widely debated in the last years (Howell, Breivik & Wilcox, 2007a,b; Bagozzi, 2007; Bollen, 2007; Wilcox, Howell & Breivik, 2008; Diamantopoulos, Riefler & Roth, 2008). While some authors show great interest towards the use of formative constructs, others question their logical consistency and suggest to definitely avoid them, in favour of reflective measurement models. At the heart of the criticism against formative constructs there is the asserted neat distinction between the existence and the measurement of a latent variable. According to Howell et al. (2007a), latent constructs are not inherently reflective or formative; they simply exist in itself, apart from the way they are measured. In particular, the existence of a latent construct would be implied by the fact that its manifest indicators vary as a function of that construct, accordingly any latent variable may, at least in principle, be measured reflectively. In addition, other practical reasons would suggest avoiding formative constructs. The manifest variables forming the latent construct need not share the same antecedents and consequences (Jarvis, MacKenzie, & Podsakoff, 2003; Howell, Breivik & Wilcox, 2007b) and need not be correlated, so that it is not clear how they can be melt together into a single latent construct. Secondly, the practical estimation of formative models requires the formative constructs to be embedded in a larger model, with a reflective component (Howell et al., 2007a). This way, formative constructs are not simply a composite of their measures, but are those composites that best predict the dependent variables in the analysis (Heise, 1972, p. 160). This would heavily expose formative models to the risk of interpretational confounding, that is to the divergence of the empirical meaning of the formative latent variables (as it comes out of the estimation procedure) from its theoretical meaning (as conceptualized without any reference to the way the latent constructs are measured). As a result, formative constructs would be scarcely useful in order to accumulate knowledge on a specific subject (Howell et al., 2007a). They would not correspond to real concepts and would be useful only as prediction tools, in a constructivist or operationalist perspective.

The position of Howell et al. (2007a) is very radical and, even if they raise some pertinent problems, we do not agree with both their essential opposition to formative models and with the asserted supremacy of reflective constructs. Reflective models are not immune from many practical problems and are not so effective in accumulating knowledge, as the indeterminacy problem of the Lisrel approach or the difficulty in interpreting PLS-PM scores make clear. At the same time, the use of formative models is spreading and suggesting to completely avoid their employment is conceptually forcing and not realistic in practice. To answer the issues raised by Howell et al. (2007a), it is useful to address the problem of formative and reflective latent constructs from a different point of view, shifting the focus from latent variables as stand alone entities, to latent variables as entities embedded in relational networks. In Howell et al. (2007a), latent concepts are thought as existing in themselves, prior to and independently from any measurement process or any relation to other latent or manifest concepts. This seems to us quite
an abstract point of view. *De facto*, latent concepts cannot be conceived outside of conceptual networks, conceptual domains or contexts and pre-existing bodies of knowledge; they are naturally embedded in them and are interrelated, explicitly or not, with other concepts to the extent that their meaning depends upon such inter-relations. This is true, in particular, for intangible concepts, like those encountered in socio-economic studies, in fact

“[…] it is primarily this phenomenon of conceptual embedding that gives rise to the intuitive notion that metaphysical concepts are more complex than concrete concepts, because metaphysical concepts are typically embedded in larger theories […]”

(Keller and Lehman, 1991, pp. 274–275)

Latent variables are supposed to capture latent concepts; explicitly or not, they are designed jointly with their relationships to other manifest or latent constructs. For example, reflective latent variables are defined (before than ‘measured’) by embedding them into a system of reflective relationships with a set of covaring manifest variables. Exogenous and endogenous latent variables in formative-reflective models are defined as ‘bridges’ between the manifest formative and reflective sides of the model. They are defined as nodes of a relational network, rather than as pre-existing entities. Shifting the focus to relational structures answers the main epistemological issue raised by (Howell et al., 2007a) against formative constructs. Inserting formative latent variables in a larger model with reflective components is no further a problem; it simply means that the relational structure defining the variables of interest comprises both a formative and a reflective side. Yet, we agree with Howell et al. (2007a) that models comprising formative relations can be poorly built, causing interpretational difficulties. But these problems can be ascribed to a misspecification of the relational structure defining the latent constructs of interest Bollen (2007) and can be avoided through a careful selection of the variables involved in the model.

Assuming a relational perspective on latent variables reflects on the way formative-reflective models must be addressed from a statistical point of view. The leading criterion in the extraction of exogenous and endogenous latent variables cannot reduce to simply maximizing the explanation power towards the reflective side of the model, but must take into account all the components of the relational scheme. In the next paragraphs we discuss this issue, presenting the new extraction procedure.

3. Latent variable extraction in formative-reflective models

In this section, we give a formal definition of formative-reflective models and develop the extraction procedure, with the aim of building a statistical tool consistent with the discussion developed in the previous section.

3.1. The structure of formative-reflective models

Formative-reflective models are an extension of the formative first-order, formative second-order model cited in Diamantopoulos et al. (2008), where more than one reflective endogenous latent variable is allowed. In such models, $p$ blocks of formative manifest variables (MVs) form $p$ different exogenous latent variables (LVs) which, in turn, form $q$ endogenous LVs, that are reflected by $q$ blocks of MVs (see Figure 1). Both formative and reflective blocks of MVs are usually retained
as unidimensional; this restricts the field of application of such models, but makes them easier to handle from the statistical point of view.

Analytically, a formative-reflective model with \( p \) exogenous LVs and \( q \) endogenous LVs is defined as follows. Let \( x_i, i = 1, \ldots, p \), and \( y_j, j = 1, \ldots, q \), be vectors of zero-mean manifest variables and let \( \omega_i, i = 1, \ldots, p \), be vectors of real coefficients. According to the formative-reflective scheme, each exogenous LV \( \xi_i \) is expressed as a linear combination of the MVs of the corresponding formative group:

\[
\xi_i = \omega'_i x_i, \quad i = 1, \ldots, p. \tag{1}
\]

By stacking \( \xi_1, \ldots, \xi_p \) and \( x_1, \ldots, x_p \) into the vectors \( \xi \) and \( x \) respectively, definitions (1) can be cast in the following compact form

\[
\xi = \Omega x, \tag{2}
\]

with

\[
\Omega = \begin{bmatrix}
\omega'_1 & 0' & \cdots & 0' \\
0' & \omega'_2 & \cdots & 0' \\
\vdots & \vdots & \ddots & \vdots \\
0' & 0' & \cdots & \omega'_p
\end{bmatrix}, \quad \xi = \begin{bmatrix}
\xi_1 \\
\xi_2 \\
\vdots \\
\xi_p
\end{bmatrix}, \quad x = \begin{bmatrix}
x_1 \\
x_2 \\
\vdots \\
x_p
\end{bmatrix}.
\]

In an analogous way, the vector \( y_j \) of MVs of the \( j \)-th reflective block are assumed to be built as sums of a rescaled common scalar endogenous LV \( \eta_j \) plus a residual \( \epsilon_j \),

\[
y_j = \lambda_j \eta_j + \epsilon_j, \quad j = 1, \ldots, q,
\]

where \( \lambda_j, j = 1, \ldots, q \), are real vectors. Again, stacking \( y_1, \ldots, y_q \) and \( \epsilon_1, \ldots, \epsilon_q \) into the vectors \( y \) and \( \epsilon \), respectively, we get

\[
y = \Lambda \eta + \epsilon, \tag{3}
\]

with

\[
\Lambda = \begin{bmatrix}
\lambda_1 & 0 & \cdots & 0 \\
0 & \lambda_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_q
\end{bmatrix}, \quad \eta = \begin{bmatrix}
\eta_1 \\
\eta_2 \\
\vdots \\
\eta_q
\end{bmatrix}, \quad \epsilon = \begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\vdots \\
\epsilon_q
\end{bmatrix}, \quad y = \begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_q
\end{bmatrix}.
\]
Finally, the endogenous LVs stacked in vector $\eta$ are built as linear combinations of the exogenous LVs stacked in vector $\xi$, namely

$$\eta = \Gamma \xi,$$ (4)

where $\Gamma$ is a conformable matrix of real coefficients. By putting (2), (3) and (4) together we obtain the final model, linking the formative MVs and the reflective MVs, via the constrained latent structure expressed by the matrices $\Lambda$, $\Gamma$, $\Omega$:

$$y = \Lambda \Gamma \Omega x + \varepsilon.$$

3.2. Latent variables extraction

In a formative-reflective scheme, exogenous LVs play a double role. On one hand, each of them should summarize effectively its own formative block; on the other hand, as a whole, they should indirectly predict (via the set of endogenous LVs) the variables in the reflective groups. The extraction of variables $\xi_1, \ldots, \xi_p$ requires a compromise between these two goals, that can be cast into a least squares optimization problem.

The first goal would be achieved extracting exogenous LVs through the minimization of the following loss function:

$$L_x(\Pi, \Omega) = \frac{1}{p} \sum_{i=1}^{p} \frac{\text{Tr}\{E[(x_i - \pi_i \omega_i' x_i)(x_i - \pi_i \omega_i' x_i)']\}}{\text{Tr}\{E[x_i x_i']\}},$$ (5)

where $\pi_i$, $i = 1, \ldots, p$, are vectors of regression coefficients of $x_i$ on $\omega_i' x_i$, and

$$\Pi = \begin{bmatrix} \pi_1 & 0 & \ldots & 0 \\ 0 & \pi_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \pi_p \end{bmatrix}.$$

Notice that $L_x$ is a weighted sum of the $p$ separated least squares loss functions $\text{Tr}\{E[(x_i - \pi_i \omega_i' x_i)(x_i - \pi_i \omega_i' x_i)']\}$, which are usually referred to as residual variances. We divide by the trace of the covariance matrix of each $x_i$ to avoid the effect of the different dimension and variability in the blocks. As a result $L_x$ takes value in the interval $[0, 1]$. As well-known (e.g. Rao, 1972, p.593) this loss function is minimised by taking each $\omega_i$ as first eigenvector of $E[x_i x_i']$.

The second goal would be achieved by extracting the exogenous (and thus the endogenous) LVs through the minimization of the following loss function:

$$L_y(\Omega, \Gamma, \Lambda) = \frac{1}{q} \sum_{j=1}^{q} \frac{\text{Tr}\{E[(y_j - \lambda_j \gamma_j' \Omega x)(y_j - \lambda_j \gamma_j' \Omega x)']\}}{\text{Tr}\{E[y_j y_j']\}},$$ (6)

where $\gamma_j$ is the $j$-th row of matrix $\Gamma$ and the vectors $\lambda_j$ are regression coefficients of $y_j$ on $\gamma_j' \Omega x$. Again, $L_y$ may be seen as an average of normalized residual variances. If the matrix $\Omega$ were given, and so the latent variables $\xi = \Omega x$, then $L_y$ would be minimised by a rank-one reduced rank regression of each $y_j$ on $\xi$ (Anderson, 1951; Reinsel, 2006).
Then, the problem of extracting the exogenous and endogenous LVs taking into account both competitive goals can be solved by minimizing the following global loss function

\[ L^{(a)}(\Pi, \Omega, \Gamma, \Lambda) = (1 - a) L_x(\Pi, \Omega) + a L_y(\Omega, \Gamma, \Lambda) \]  

(7)

where \( a \in [0, 1] \) determines the relative weight given to each goal.

When \( a = 0 \), \( \omega_i \) is just the first eigenvector of \( \mathbb{E}[x_i x_i'] \), \( \xi_i \) is the first principal component of the variables in the \( i \)-th formative block and \( y_j = \lambda_j \gamma_j' \xi + \varepsilon_j \) is an ordinary rank-one reduced rank regression of \( y_j \) on \( \xi \). On the contrary, when \( a = 1 \) the latent variables \( \xi_i \)'s are built as the linear combinations of the respective \( x_i \) that best fit, via the endogenous latent variables, the vector \( y \), and the whole problem reduces to a multivariate regression with many constrains, implied by the particular form of the regression coefficient matrix \( B = \Lambda \Gamma \Omega \).

The existence of a solution to the optimization problem, for any \( a \in [0, 1] \), follows from the fact that \( L^{(a)} \) is a convex combination of least squares loss functions. When \( a = 0 \), the solution is essentially unique since the exogenous latent variables are just the first principal components of the corresponding formative blocks. For \( a \in (0, 1] \) unicity cannot be guaranteed in general, but our conjecture is that cases of non-unicity are rather rare, artificial or pathological and our experience working with real data confirms this.

Due to constraints in the matrices \( \Omega \) and \( \Lambda \), the minimization of \( L^{(a)} \) with respect to all the parameters is nonlinear and the optimal solution must be found numerically. Despite the presence of many parameters in the loss function, the majority of them can be concentrated out, and the numerical minimization can be carried out only with respect to the free parameters in \( \Omega \), as shown below.

Consider the loss function \( L_x \) and let, for two generic vectors \( w \) and \( z \), \( \Sigma_{wz} = \mathbb{E}[(w - \mathbb{E}[w])(z - \mathbb{E}[z])'] \) be their covariance matrix. For any given \( \Omega \) (and thus \( \xi \)) the matrix \( \Pi \) that minimizes \( L_x \) is obtained by collecting the least-squares regression coefficient vectors \( \pi_i \) of \( x_i \) on \( \xi_i \). The covariance matrix of the regression residuals is given by

\[ S_x^{(i)}(\Omega) = \Sigma_{x_i x_i} - \Sigma_{x_i x_i} \Sigma_{\xi_i \xi_i}^{-1} \Sigma_{\xi_i x_i} = \Sigma_{x_i x_i} - \Sigma_{x_i x_i} \omega_i (\omega_i' \Sigma_{x_i x_i} \omega_i)^{-1} \omega_i' \Sigma_{x_i x_i}, \]

so that for any given \( \Omega \)

\[
\min_{\Pi} L_x(\Pi, \Omega) = \frac{1}{p} \sum_{i=1}^{p} \frac{\text{Tr}[S_x^{(i)}(\Omega)]}{\text{Tr}[\Sigma_{x_i x_i}]}. \]

As for the \( L_y \) function, the numerator of the generic summand in equation (6) may be rewritten as

\[ \text{Tr}[\mathbb{E}[(y_j - \lambda_j \gamma_j' \xi)(y_j - \lambda_j \gamma_j' \xi)']]. \]

(8)

So, given the vector \( \xi \), the minimization of \( L_y \) is equivalent to the estimation of the \( q \) independent reduced-rank regressions

\[ y_j = \lambda_j \gamma_j' \xi + \varepsilon_j \quad j = 1, \ldots, q, \]
and equation (8) is minimized by extracting the first pair of canonical variables between $y_j$ and $\xi$ and then regressing $y_j$ on the $\xi$-side canonical variable. In formulas, $\hat{\gamma}_j$ is given by the first eigenvector of the following matrix (see, for instance, Kettenring, 2006, equation 6)

$$M = \Sigma_{y_jy_j}^{-1} \Sigma_{y_j\xi} \Sigma_{\xi\xi}^{-1} \Sigma_{\xi y_j},$$

while

$$\hat{\lambda}_j = \frac{\Sigma_{y_j\eta_j}}{\sigma^2_{\eta_j}} = \frac{\Sigma_{y_j\xi} \hat{\gamma}_j}{\gamma_j \Sigma_{\xi\xi} \hat{\gamma}_j}.$$

The covariance matrix of the residuals of each of these (rank one) reduced rank regressions is given by

$$S^{(j)}(\Omega) = \Sigma_{y_jy_j} - \Sigma_{y_j\xi} \hat{\gamma}_j (\gamma_j' \Sigma_{\xi\xi} \hat{\gamma}_j)^{-1} \hat{\gamma}_j / \Sigma_{\xi y_j},$$

where the dependence of $\hat{\gamma}_j$ on $\Omega$ has been suppressed to keep the notation light. The vectors $\lambda_j$ do not appear in $S^{(j)}$ since, given $\eta_j = \gamma_j \xi$, they are just regression coefficients that can be concentrated out. Thus, also $\Gamma$ has been concentrated out and the optimization reduces to taking the minimum of

$$L^{(a)}(\Omega) = \frac{(1 - \alpha)}{p} \sum_{i=1}^p \frac{\text{Tr}[S^{(i)}(\Omega)]}{\text{Tr}[\Sigma_{x_i x_i}]} + \frac{\alpha}{q} \sum_{j=1}^q \frac{\text{Tr}[S^{(j)}(\Omega)]}{\text{Tr}[\Sigma_{y_j y_j}]}$$

with respect to $\omega_1, \ldots, \omega_p$ only.

Notice that the complement to 1 of the loss functions $L_x$, $L_y$ and $L^{(a)}$ are related to the coefficients of determination. In particular, since

$$R^2_x = 1 - \frac{\text{Tr}[S^{(x)}(\Omega)]}{\text{Tr}[\Sigma_{x_i x_i}]}$$

$$R^2_y = 1 - \frac{\text{Tr}[S^{(y)}(\Omega)]}{\text{Tr}[\Sigma_{y_j y_j}]}$$

then $R^2_x = 1 - L_x$ is just the mean of the $p$ coefficients of determination for the $x$-side, $R^2_y = 1 - L_y$ is the mean of the $q$ coefficients of determination for the $y$-side and $R^2(\alpha) = 1 - L^{(a)}$ is the $\alpha$-weighted average of $R^2_x$ and $R^2_y$.

To perform the minimization of $L^{(a)}(\Omega)$, the BFGS algorithm with numerical derivatives has been employed, taking the first eigenvector of $\Sigma_{x_i x_i}$ as starting value for $\omega_i$. Since the $\omega_i$ vectors are defined up to a scale factor, before the numerical optimization can be carried out, it is necessary to fix either one of the elements of each $\omega_i$ (e.g. set the first element of each $\omega_i$ to 1) or the norms of $\omega_i$ (e.g. $\|\omega_i\| = 1$).

3.3. The case of multidimensional formative blocks

Up to now, we have considered unidimensional formative blocks only, but in principle also multidimensional blocks could be of concern. The algorithm discussed above can be easily extended to such a case, allowing for the extraction of $k_i$ latent variables $\xi_{i,1} \ldots \xi_{i,k_i}$ for each $x$-group $i$, that is defining

$$\xi_i = \Omega_i x_i,$$
where $\Omega_i$ is a $k_i \times m_i$ matrix and $m_i$ is the number of variables in group $i$. Notice that $\Omega$ is now defined by

$$
\Omega = \begin{bmatrix}
\Omega_1 & 0 & \ldots & 0 \\
0 & \Omega_2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & \Omega_p
\end{bmatrix}
$$

The covariance matrix in each summand of the $x$-side loss function changes into

$$
S_{x}^{(i)}(\Omega) = \Sigma_{x,x_i} - \Sigma_{x,x_i} \Omega_i' (\Omega_i \Sigma_{x,x_i} \Omega_i')^{-1} \Omega_i \Sigma_{x,x_i},
$$

while all the other formulas remain valid.

The vector of latent variables $\xi_i$ is now defined up to multiplication by a $k_i \times k_i$ full-rank matrix. In order to solve this indeterminacy problem, $\xi_{i,1}, \ldots, \xi_{i,k_i}$ are orthonormalized, choosing the set of $k_i$ normalized principal components as the orthogonal basis for the linear space spanned by the $k_i$ components of $\xi_i$.

From an operational point of view, this can be achieved in two steps.

1. For each $i$ in $1, \ldots, p$, fix $k_i^2$ elements in $\Omega_i$ using values from an arbitrary $k_i \times k_i$ full rank matrix and minimize the loss function $L^{(a)}(\Omega)$ with respect to the free elements of $\Omega_1, \ldots, \Omega_p$. This allows the extraction of the variables in vector $\tilde{\xi}_i = \tilde{\Omega}_i x_i$, which span the space we are interested in.

2. In order to extract the variables spanning the space extracted in step 1, but following a principal component rationale (orthogonality and maximal explained variance), we carry out a canonical correlation analysis of $x_i$ with $\tilde{\xi}_i$.

For each $i$, we keep all the $k_i \tilde{\xi}_i$-side canonical variables as final solutions $\hat{\xi}_i$.

4. Simple examples

In order to illustrate the use of the proposed procedure and to show what the application of our technique on a grid of $\alpha$ values can reveal about the model, we now present four simple examples based on artificial correlation structures. We first concentrate on a simple model with two groups of $x$-variables each containing just two manifest variables and with one group of $y$-variables also with two manifest variables (see Figure 2). In Example 4 we will slightly modify the model, inserting one more manifest variable in the first $x$-group.

![Figure 2: Path diagram of the model used in Examples 1-3.](image)

The analysis of the performance of the fitted models will be based on the following quantities:
• $R^2_x$, coefficient of determination for the $x$-variables;
• $R^2_y$, coefficient of determination for the $y$-variables;
• $R^2_{xy} = 0.5R^2_x + 0.5R^2_y$;
• $R^2_\alpha = (1 - \alpha)R^2_x + \alpha R^2_y$;
• $\text{Corr}(\xi_i(\alpha), \xi_i(0))$, correlation of each $\xi_i$ variable extracted at a given $\alpha$ with the same variable extracted as the first principal component of its manifest $x$-group (as when $\alpha = 0$).

The last quantity is useful for assessing how much the LVs $\xi_i$ rotate, when the analyst’s goal gradually shifts from summarising the $x$-blocks to fitting the $y$-variables, as $\alpha$ varies from 0 to 1. In practice, such a correlation is used to check whether the formative and the reflective sides of the model are jointly specified in a consistent way.

**Example 1.** Consider the following covariance (correlation) matrix $\Sigma$:

\[
\Sigma = \begin{bmatrix}
  x_{11} & x_{12} & x_{21} & x_{22} & y_{11} & y_{12} \\
  x_{11} & 1.0 & 0.8 & 0.7 & 0.8 & 0.8 \\
  x_{12} & 0.8 & 1.0 & 0.7 & 0.8 & 0.8 \\
  x_{21} & 0.7 & 0.7 & 1.0 & 0.8 & 0.8 \\
  x_{22} & 0.7 & 0.7 & 0.8 & 1.0 & 0.8 \\
  y_{11} & 0.8 & 0.8 & 0.8 & 0.8 & 1.0 \\
  y_{12} & 0.8 & 0.8 & 0.8 & 0.8 & 1.0 
\end{bmatrix}
\]

Here, correlations within $x$-groups are very high, and $x$-variables are also highly correlated with $y$-variables. Applying the estimation algorithm for a grid of $\alpha$ values in $[0, 1]$, we get the results depicted in Figure 3.

![Figure 3: Determination coefficients and rotation of the $\xi$'s as functions of $\alpha$ for Example 1.](image)

The coefficients $R^2_x$ and $R^2_y$ are rather large and do not depend on $\alpha$, as a consequence $R^2_\alpha$ changes only through the weight $\alpha$. The correlation of the extracted $\xi_i$ with the respective principal components is constantly equal to one irrespective of the value of $\alpha$. Given the structure of the covariance matrix, this means that the latent variables extracted by the algorithm are stable as $\alpha$ varies, showing
that the formative-reflective scheme is well-specified and that the formative and
the reflective sides of the model may consistently be connected through the causal
relationships specified in the model.

Example 2. Let the covariance matrix of the manifest variables be

\[
\Sigma = \begin{pmatrix}
1.0 & 0.2 & 0.8 & 0.0 & 0.9 & 0.9 \\
0.2 & 1.0 & 0.0 & 0.0 & 0.2 & 0.2 \\
0.8 & 0.0 & 1.0 & 0.2 & 0.8 & 0.8 \\
0.0 & 0.0 & 0.2 & 1.0 & 0.1 & 0.1 \\
0.9 & 0.2 & 0.8 & 0.1 & 1.0 & 0.8 \\
0.9 & 0.2 & 0.8 & 0.1 & 0.8 & 1.0
\end{pmatrix}
\] (10)

Here, the \(x\)-variables are barely correlated within their groups, but the first \(x\)-variable of each group is highly correlated with both variables in the \(y\)-group. The \(x\)-variables are scarcely collinear, but can account for a significant part of the \(y\)-side variance. As depicted in Figure 4, \(R^2_x\) is relatively small for every \(\alpha\). On the contrary, as \(\alpha\) increases, \(R^2_y\) grows significantly. This means that a formative-reflective scheme does not seem appropriate in this case. The explicative power of the formative blocks could be better exploited through a reduced-rank regression of the reflective manifest variables on the variables comprised into the formative blocks. As the second panel of Figure 4 shows, the \(\xi_i\)'s rotate with \(\alpha\), but the rotation is relatively mild.

Example 3. Let us consider the covariance matrix

\[
\begin{pmatrix}
x_{11} & x_{12} & x_{21} & x_{22} & y_{11} & y_{12} \\
x_{12} & 0.9 & 0.0 & 0.0 & 0.1 & -0.1 \\
x_{21} & 0.0 & 1.0 & 0.0 & 0.0 & -0.1 \\
x_{22} & 0.0 & 0.0 & 1.0 & 0.8 & 0.2 & 0.0 \\
y_{11} & 0.1 & 0.0 & 0.2 & 1.0 & 0.1 & 0.0 \\
y_{12} & -0.1 & -0.1 & 0.0 & 0.5 & 1.0
\end{pmatrix}
\] (11)
In this example, variables within each \( x \)-group are highly correlated, but they are scarcely correlated with the \( y \)-variables. From Figure 5, \( R_x^2 \) appears to remain very close to zero, while \( R_y^2 \) is generally large (the \( x \) are rather collinear in each group), dropping for values of \( \alpha \) very close or equal to one, for which the reflective side of the model tends to prevail. This dramatic reduction of \( R_x^2 \) comes with a strong rotation of \( \xi_1 \). In this case the \( \xi_i \)'s are well defined, but they are clearly not able to take account of the variance of the \( y \)-variables.

**Example 4.** In this last example we slightly modify the structure depicted in Figure 2, by introducing a third variable in the first \( x \)-group. Let the covariance matrix be

\[
\Sigma = \begin{bmatrix}
X_{11} & X_{12} & X_{13} & X_{21} & X_{22} & Y_{11} & Y_{12} \\
X_{11} & 1.0 & 0.8 & 0.1 & 0.0 & 0.0 & 0.0 \\
X_{12} & 0.8 & 1.0 & 0.1 & 0.0 & 0.0 & 0.0 \\
X_{13} & 0.1 & 0.1 & 1.0 & 0.0 & 0.0 & 0.5 \\
X_{21} & 0.0 & 0.0 & 0.0 & 1.0 & 0.2 & 0.1 & 0.5 \\
X_{22} & 0.0 & 0.0 & 0.0 & 0.2 & 1.0 & 0.1 & 0.1 \\
Y_{11} & 0.0 & 0.0 & 0.7 & 0.1 & 0.1 & 1.0 & 0.5 \\
Y_{12} & 0.0 & 0.0 & 0.5 & 0.5 & 0.1 & 0.5 & 1.0 \\
\end{bmatrix} \tag{12}
\]

Here, the first \( x \)-group is not well designed. Indeed, there are two variables that are highly correlated with each other and a third that is almost orthogonal to both of them. However, the first two variables of this group are uncorrelated with both \( y \)-variables, while the third is positively correlated with both of them.

As Figure 6 illustrates, in this case the extraction of \( \xi_1 \) changes radically according to \( \alpha \). For \( \alpha \) close to zero, the first two elements of the weight vector \( \omega_1 \) are much higher than the third one, since the first principal component of this group is taking account of the collinearity of \( x_{11} \) and \( x_{12} \). For values of \( \alpha \) close to 1, \( \xi_1 \) is approximately proportional to \( x_{13} \) since the model is fitting the \( y \)-variables as well as possible. In this case, the analyst should restructure the first \( x \)-group, maybe isolating \( x_{13} \) in a third group, possibly including other meaningful variables.
5. A real world application

In order to illustrate our method and compare it with the popular PLS-PM approach, we replicate the analysis of Bruhn, Georgi & Hadwick (2007), who make the correlation matrix of their variables publicly available (Table 3 of the cited paper). The objective of Bruhn et al. (2007) is evaluating “how intensively firms orient their customer management toward customer value and equity.” In particular, they focus on the concept of customer equity management (CEM) defined as “all [those business] activities that aim explicitly to maximize customer equity.” The data for their study are obtained from the submission of a questionnaire to managers of Swiss and German companies in six industries with high CEM importance: airlines, banking, mail order businesses, telecommunications, tourism and utilities.

We have first run our extraction procedure setting $\alpha = 1$ (y-side optimization only) in order to comply with the approach chosen by the authors, whose goal was mainly to explain the reflective side of the model. The results are summarized in the path diagram of Figure 7 and can be compared with those in Bruhn et al. (2007) which are reported in parenthesis. Notice that Bruhn et al. (2007) do not report the $R^2_x$ and $R^2_y$ we show in Figure 7 as PLS-PM software packages usually report only the $R^2$ relative to the projection of endogenous LVs on exogenous LVs. Since in our model the endogenous LVs lay in the space spanned by the exogenous LVs, the corresponding $R^2$ for our procedure would be 1, making the comparison uninteresting. The $R^2_y$ of Bruhn et al.’s estimates is easily obtained as mean of the squared correlations of exogenous LVs and MVs, while their $R^2_x$ is derived as mean of the products of the reported $R^2$ (for the projection of endogenous LV on exogenous LVs) with the squared correlations of endogenous LV and MVs.

As it can be directly checked:

1. Differently from Bruhn et al.’s output, the exogenous MVs and the exogenous LV are highly correlated, that is the latent variables CE analysis, CE Strategy and CE Action actually summarize their respective formative blocks. This is also revealed by their overall fit, measured by $R^2_y$, which equals 0.63, much greater than the corresponding fit obtained by Bruhn et al. (0.17).

2. The correlation between the endogenous LV ($\hat{\eta}_{LS}$) and each of the exogenous LVs is also much greater than in Bruhn et al.’s output. Differently from PLS-PM, our procedure builds the endogenous LV as a linear combination of the
three exogenous LVs and its meaning is clearly interpretable in terms of the formative side of the model.

3. The endogenous LV is nonetheless highly correlated with each of the three y-variables. Clearly, PLS-PM correlations are higher, but at the cost of expressing the endogenous LV (\( \hat{\eta}_{PLS} \)) as a linear combination of the y-variables, inconsistently with the logical structure of the model and with low and poorly interpretable correlations with the latent variables CE analysis, CE Strategy and CE Action.

4. The predictive power of the endogenous MVs achieved by our procedure is given by \( R^2_y = 0.44 \), whereas the same coefficient for PLS-PM is 0.43. This fact reveals the non-optimality of PLS-PM, as the scarcer representativeness of its exogenous LVs is not compensated by a better fit to the endogenous MVs.

The goal of a causal model is to identify which causes affect more intensively the outcomes and to suggest how to act on the former, to obtain the desired effects. Our procedure does accomplish this task: the formative and the reflective MVs are connected in a balanced and understandable way through the endogenous part of the model. The results do not suffer of interpretational confounding problems, since both the x- and y-manifest sides are jointly taken into account when extracting the latent variables. The causal flows from xs to ys are neat. PLS-PM is not designed for formative-reflective schemes and the outputs of Bruhn et al. (2007) reveal the drawbacks of applying it to this kind of structural model, as discussed above. The results obtained by PLS-PM are exposed to the critics of Howell et al. (2007a) and are even scarcely useful in a decision-making perspective, since no clear causal paths are revealed, between the manifest formative and reflective sides of the model, albeit they exist, as the output of our procedure show. Our results are even reinforced computing the coefficients of determination and rotation defined in the previous sections for a grid of \( \alpha \)-values, to show that the results of our procedure are very stable as \( \alpha \) varies. As shown in Figure 8, the value of \( \alpha \) does not
affect much either $R^2_x$ or $R^2_y$, indicating that the three exogenous IVs are contemporaneously trustworthy representatives of the $x$-variables and good predictors of the $y$-variables. This is also confirmed by the right panel of Figure 8, where the $\xi$ variables undergo only mild rotations as $\alpha$ varies in $[0, 1]$: the maximum drop in the correlation of $\xi_i(\alpha)$ with $\xi_i(0)$, for $i = 1, 2, 3$, is in fact from 1.00 to 0.91. This means that the exogenous IVs are well-defined and stable over the whole range of $\alpha$ values, not suffering of interpretational confounding problems.

6. Conclusion

In this paper, we have proposed a new procedure for extracting latent variables in a formative-reflective model, which proves to be a valuable alternative to classical tools such as Lisrel and PLS-PM. The main feature of the procedure is that the global optimum criterion takes into account both the capability of the exogenous latent variables to summarize their manifest formative blocks and the ability, via the endogenous latent variables, to explain the reflective manifest blocks. The procedure can also be employed to investigate the relational consistency of the model, checking how the extracted variables change as more importance is given to the formative side or to the reflective side of the model. Our extraction procedure overcomes the limitations of the classical tools, in that latent scores are uniquely determined and the causal structure of the model is fully respected.

A complete set of software routines implementing the procedure has been developed and made freely available. The effectiveness of the procedure has been shown and discussed through some simulated examples and a real data application, where, in particular, the advantages over PLS-PM emerge quite clearly. The procedure can be extended in many directions and there is room for further research, mainly pertaining to (i) taking into account errors in equations (Vittadini et al., 2007), which at present are not considered in the extraction algorithm, and (ii) the inferential extension of the procedure. Notwithstanding these possible developments, the proposed procedure already provides a valuable tool for managing formative-reflective schemes, making the use of these important models more robust and effective from a methodological and a statistical point of view.
References


