1 The partial robust M-approach

Sven Serneels[1], Christophe Croux[2], Peter Filzmoser[3], and Pierre J. Van Espen[1]

[1] Department of Chemistry
University of Antwerp, 2610 Antwerp, Belgium
[2] Department of Applied Economics,
KULeuven, 3000 Leuven, Belgium
[3] Department of Statistics and Probability Theory,
Technical University of Vienna, 1040 Wien, Austria.

Abstract. The PLS approach is a widely used technique to estimate path models relating various blocks of variables measured from the same population. It is frequently applied in the social sciences and in economics. In this type of applications, deviations from normality and outliers may occur, leading to an efficiency loss or even biased results. In the current paper, a robust path model estimation technique is being proposed, the partial robust M (PRM) approach. In an example its benefits are illustrated.

1.1 Introduction

Consider the situation where one disposes of j blocks of observable variables, each of which one supposes to be the effect of a sole unobservable, *latent* variable. Furthermore, structural relations between the latent variables of the different groups are assumed to exist. Different techniques to estimate these latent variables as well as the relations between them, have been proposed in literature.

On the one hand, one can use maximum likelihood techniques such as LISREL (Jöreskog and Sörbom 1979), where rigid model assumptions concerning multinormality have to be verified. If one desires less rigid assumptions, so-called *soft* modelling might prove a viable alternative. The most successful approach to soft modelling of the problem described before, is the so-called *PLS approach* (Wold 1982), which moreover gives the benefit of estimating the latent variables at the level of the individual cases, in contrast to LISREL. The PLS approach is also known as *PLS path modelling* or as *PLS structural equation modelling*. A myriad of applications of the PLS approach have been reported in literature, the most salient one probably being the European Customer Satisfaction Index (Tenenhaus et al. 2005).

The simplest path model one can consider is a path model relating a block of variables $\mathbf{x}$ to a univariate variable y, through a latent variable ξ. Model estimates for this setting can also be used for prediction of y. Hence the PLS approach, relating two blocks of variables to each other over a single

latent variable (which may be a vector variable), can be used as a *regression* technique.

The PLS estimator can be seen as a *partial* version of the least squares estimator. The latter has properties of optimality at the normal model. However, at models differing from the normal model, other estimators such as the M-estimator may have better properties (Huber 1981). Especially for heavy-tailed distributions such as the Cauchy distribution or the ε-contaminated normal distribution, partial versions of robust estimators may be expected to out-perform PLS. Hence, in a recent paper we have proposed the *partial robust M-regression* estimator (Serneels et al. 2005). Simulations have corroborated the aforementioned assumptions. As the PLS regression estimator is very sensitive to outliers and extreme values, the same holds for the PLS approach as a whole, since a PLS regression is carried out at each iteration.

In the current paper, we propose a robust version of the PLS approach based on the robust M-estimator, which will be called the *Partial Robust M-approach*. An example will show the beneficial properties of the novel approach introduced here.

1.2 The model and the partial robust M-approach

Before we can proceed with the description of the partial robust M-approach, we first provide a brief introduction to the PLS approach. More elaborate introductions can be found in the works of Tenenhaus (1999) and Chin and Newsted (1999). Suppose one disposes of j blocks of centred observable variables $\mathbf{x}_i = x_{i1}, \cdots, x_{ik_i} (i \in 1, 2, \cdots, j)$, where k_i denotes the number of variables in block i. These variables are referred to as the *manifest* variables. Each of these groups of variables can be considered to be essentially univariate: they are the observable counterpart of a single latent variable ξ_i. Manifest and latent variables are related to each other by the linear model $(h \in 1, \cdots, k_i)$:

$$x_{ih} = \varpi_{ih}\xi_i + \varepsilon_{ih}. \tag{1.1}$$

It is supposed that the random error term ε_{ih} has zero expectation and is non-correlated to the latent variable. The studied phenomenon is assumed to have been generated by structural relations between the latent variables

$$\xi_i = \sum_q \beta_{iq}\xi_q + \phi_i, \tag{1.2}$$

where it is assumed that the random error term ϕ_i has zero expectation and is not correlated to the latent variable ξ_i.

In practice, the latent variables are estimated as linear combinations y_i of the manifest variables x_{ih}:

$$y_i = \sum_h w_{ih}x_{ih} = \mathbf{w}_i^T\mathbf{x}_i \tag{1.3}$$

The vectors $\mathbf{w}_i$ are called the *weights*. However, due to the structural relations (1.2), another estimate z_i of ξ_i is given by:

$$z_i \propto \sum_{q \neq i} c_{qi} y_q. \tag{1.4}$$

The sign $\propto$ indicates that the variable on the left hand side of the Equation sign is the standardized version of the expression on the right hand side.

Several estimation schemes exist. In this paper we will limit ourselves to the so-called *centroïd* scheme, as this is the only scheme which will be used in the following section (a motivation thereto can be found in Tenenhaus, 1998). In the centroid scheme, it is necessary for the operator to specify the expected sign $c_{iq} = \text{sgn}(\text{corr}(\xi_i, \xi_q))$, where c_{iq} is set to zero if the latent variables considered are not expected to be correlated.

In the original work by H. Wold, two modes for estimation of the weights were proposed. Here we will limit our discussion to what Wold referred to as "mode A", which corresponds to the definition of the weights in PLS regression:

$$\mathbf{w}_i = \text{cov}\,(\mathbf{x}_i, z_i) \tag{1.5}$$

This leads to the following condition of stationarity:

$$y_i \propto \mathbf{x}_i^T \mathbf{x}_i \sum_{q \neq i} c_{qi} y_q \tag{1.6}$$

From Equation (1.6) it can be seen that the estimates for ξ_i can be obtained iteratively, starting from an initial guess y_i. It can also be seen from Equation (1.5) that in each iteration, the computation of the new values for y_i can be done by computing the first component of a PLS regression of z_i on $\mathbf{x}_i$.

A robustification of the PLS approach is now straightforward. The same iterative estimation scheme is being maintained, albeit at each step the respective PLS regressions are replaced by partial robust M-regressions (Serneels et al. 2005). Partial robust M-regression is an extension of robust M-regression to the latent variable multivariate regression scheme; in this context it has been proven to be superior to PLS if the data come from a non-normal distribution such as a Cauchy or a Laplace distribution.

It has been shown that the partial robust M-regression estimator can be implemented as an iteratively re-weighted PLS algorithm (Serneels et al. 2005), where the weights correct for both leverage and vertical outlyingness. A good robust starting value for the algorithm has been described. The use of an iterative re-weighting algorithm makes the method very fast in the computational sense.

1.3 Example: economical inequality leads to political instability

In this section we will study a data set first published by Russett (1964). It has been analyzed by PLS and PLS path modelling by Tenenhaus (1998, 1999). In the data set, five variables which were at the time thought to be representative of a country's economical situation, were included. Their relation to seven variables which correspond to political (in)stability, was studied. It has been shown that some data pre-processing was necessary in order to obtain interpretable results. In the current paper, we will not further discuss the data pre-processing, but we will assume that the variables have been pre-processed as has been described by Tenenhaus (1999). The same pre-processing has been used for the classical and robust estimation. Furthermore, 3 observations out of 45 contained missing data. These observations have been left out in the results obtained here.

The first block of variables, which correspond to the countries' economical situation, in fact consists of two blocks. The first block, comprising the first three manifest variables, are variables which describe the (in)equality in terms of the possession of land fit for agriculture. The second block of manifest variables, consisting of the remaining two variables describing a country's economical situation, correspond to the degree of industrialization in the respective country.

Hence, Tenenhaus (1999) proposed a path model, where it is assumed that each of the blocks has been generated by a single latent variable, i.e. the agricultural inequality (ξ_1), the degree of industrialization (ξ_2) and political instability (ξ_3). It is assumed that the agricultural inequality leads to political instability, whereas industrialization does not. Hence, we have obtained the coefficients c_{iq} from Equation (1.6): $c_{13} = c_{31} = 1$ and $c_{23} = c_{32} = -1$. Both remaining coefficients c_{12} and c_{21} are set equal to zero.

From Equation (1.6) we see how we can build up the iterative estimation scheme. We start from an initial guess, e.g. $y_1^{(1)}$ and $y_3^{(1)}$ are the first X and Y components obtained from a PLS regression of the political variables on the agricultural variables, whereas $y_2^{(1)}$ is taken as the first $\mathbf{x}_2$ component from a PLS regression of the political variables on the industrial variables. The superscripts indicate the iteration step. Suppose that we have in the $(r-1)$-th step of the algorithm $y_i^{(r)}$ as the then best estimates of the latent variables. Then we can update them in the rth step by the following scheme, based on Equation (1.6):

- the variable $y_1^{(r+1)}$ is the first PLS component of a PLS regression of $y_3^{(r)}$ on X_1 (X_1 is a matrix consists of n observations of $\mathbf{x}_1$);
- $y_2^{(r+1)}$ is the first PLS component obtained from a PLS regression of $-y_3^{(r)}$ on X_2;
- $y_3^{(r+1)}$ is the first PLS component obtained from a PLS regression of $y_2^{(r)} - y_1^{(r)}$ on X_3.

This processus is repeated until convergence. The robust estimates reported later in this section are obtained by the same iterative procedure, albeit the estimates $y_j^{(r+1)}$ are in that case the first components of the corresponding PRM regressions.

In path modelling it is customary to represent the path model by a flowchart. Manifest variables are displayed in boxes; latent variables are displayed in circles. The arrows show the direction in which the variables influence each other. The correlation coefficients between the manifest and latent variables are shown above the respective arrows. In order to describe the relations among the latent variables, the regression coefficients d_i describing the linear relation $y_3 = d_1 y_1 + d_2 y_2$, are shown above the arrows relating the latent variables. The results obtained by Tenenhaus (1998) are shown in Figure 1.1.

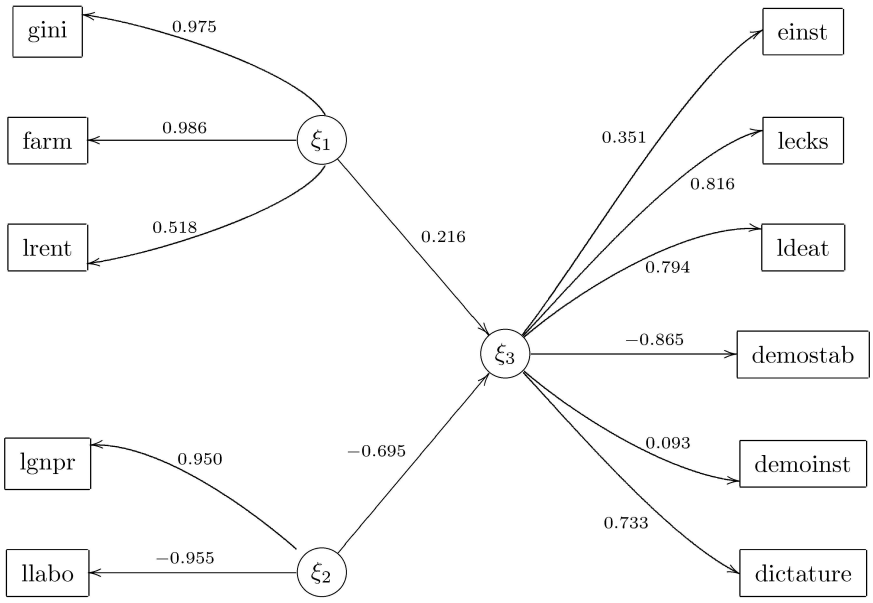

Fig. 1.1. Causality scheme estimated by Tenenhaus (1998) by dint of the PLS approach relating economical inequality and political instability.

Figure 1.1 leads Tenenhaus (1998) to the conclusion that political instability is caused rather by a lack of industrialization than by an inequality in the possession of land. However, based on economic arguments, in the original analysis by Russett it had been expected that each of the five economical variables would contribute equally to political instability.

In the data set considered here, no outliers are present in the sense that they are bad measurements which should be deleted before performing the

PLS approach. However, some influential observations are present. A good diagnostic to detect influential observations in the PLS context is the Squared Influence Diagnostic (SID) which is based on the univariate PLS influence function (Serneels *et al.* 2004). As it is a univariate test, it should be performed separately on each of the variables of X_3. A SID plot of X_2 on the variable "demostab", e.g., is plotted in Figure 1.2.

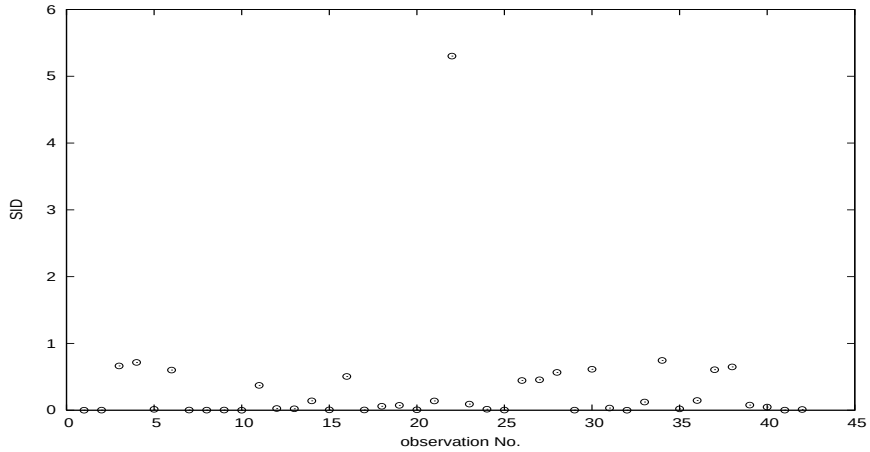

Fig. 1.2. Squared Influence Diagnostic plot for PLS1 regression of the variable "demostab" on X_2.

It unveils that the observation which corresponds to India (observation 22) is a very influential sample. This has also been signalled by Tenenhaus (1999), who notices that India is the only democracy whose level of industrialization is below the mean value. When computing the SID for other combinations of the X_i blocks and individual variables of X_3, a few other influential observations can be discerned.

The presence of some observations which are very influential on the final estimate suggests that a robust estimate might in this case suffer less from these individual observations and might be more apt to discern the general trend in the data. As a robust estimation technique, we applied the partial robust M (PRM) approach to estimate the desired quantities. The tuning constant was set to 4 (for further details see Serneels *et al.* 200x) and convergence of the partial (PRM) approach was obtained after 3 iterations, as was the case for the PLS approach. The obtained estimates are shown in Figure 1.3.

From Figure 1.3 it can be seen that the robust estimates differ somewhat from the estimates obtained by the classical PLS approach. The correlations

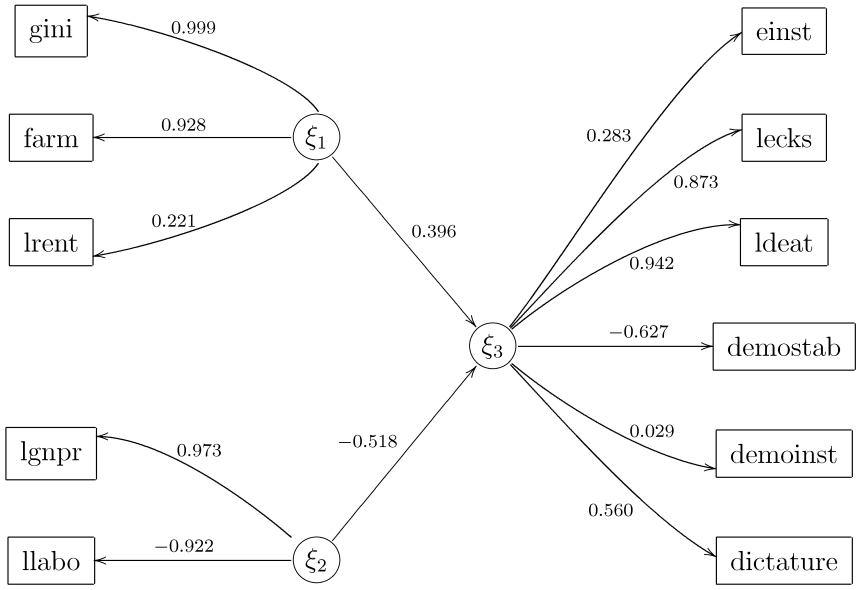

Fig. 1.3. Causality scheme estimated by dint of the PRM-approach relating economical inequality and political instability.

between the manifest variables and the latent variables show the same trend as in Figure 1.1, although some small differences may be observed: the variable "einst" is shown to be less informative whereas the variable "ldeat" is more informative to the robust model. Note that the correlations shown in the robust causality scheme (Figure 1.3) are Spearman correlations, as the usual Pearson correlations might also yield unreliable results due to deviations from normality.

The main difference between the classical and robust estimates resides in the estimation of the latent variables and the way these are related to each other. From Tenenhaus (1999) it was decided that the latent variable corresponding to the level of industrialization (ξ_2) determines to a much greater extent the country's political instability (ξ_3) than the agricultural inequality (ξ_1) does. From the robust estimates, one observes that the latter latent variable is still more important than the former, although the difference is much smaller. One could indeed conclude that both agricultural inequality and industrialization contribute about equally to political instability.

1.4 Conclusions

The PLS approach is a technique which is widely applied to estimate path models between several blocks of variables. It is believed that the path model

unveils the general trend of the structural relations which exist between these variables.

The PLS approach is very sensitive to influential observations such as outliers. These outliers might distort the final estimate in their direction.

The PLS approach is a widely applied technique in social sciences and economics. In these fields of research, influential observations are frequently not outliers which are outlying due to bad measurement which should be removed before model estimation, but outliers often correspond to individuals which behave differently than the majority of the data. Hence, the information these observations carry should be used at the model estimation step, albeit their influence in the final estimate should be controlled. The aforementioned arguments suggest the use of a robust estimation technique for the path model. Robust M-estimators are resistant with respect to outliers, but remain highly efficient at the normal model.

In the current paper, the partial robust M-approach has been proposed as a robust estimation technique for path modelling. It is based on several steps of partial robust M-regression (Serneels *et al.* 200x). In an example it has been shown to yield improvements over the PLS approach such that it can better unveil the general trend in the path model relation, in case the data do not follow a normal model.

References

Chin, W.W. and Newsted, P.R. (1999): Structural Equation modelling analysis with small samples using partial least squares. In Hoyle, R.H. (Ed.): *Statistical strategies for small-sample research*. Sage, Thousand Oaks (CA), pp. 307-341.

Huber, P.J. (1981): *Robust Statistics*. Wiley, New York.

Jöreskog, K.G. and Sörbom, D. (1979): *Advances in factor analysis and structural Equation models*. Abt books, Cambridge.

Russett, B.M. (1964): Inequality and instability. *World Politics*, 21, 442-454.

Serneels, S., Croux, C. and Van Espen, P.J. (2004): Influence properties of partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 71, 13-20.

Serneels, S., Croux, C., Filzmoser, P. and Van Espen, P.J. (2005): Partial robust M-regression. *Chemometrics and Intelligent Laboratory systems*, 79, 55-64.

Tenenhaus, M. (1998): *La régression PLS*. Technip, Paris.

Tenenhaus, M. (1999): L'approche PLS. *Revue de Statistique Appliquée*, XLVII (2), 5-40.

Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.-M. and Lauro, C.(2005): PLS path modelling. *Computational Statistics and Data Analysis*, 48, 159-205.

Wold, H. (1982): Soft modeling: the basic design and some extensions, in: K.G. Jöreskog and H. Wold (eds.), *Systems under indirect observation, vol. 2*. North-Holland, Amsterdam, pp. 1-54.