Analytica Chimica Acta 642 (2009) 45–58

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Exploring a physico-chemical multi-array explanatory model with a new multiple covariance-based technique: Structural equation exploratory regression

X. Bry[a,*], T. Verron[b], P. Cazes[c]

[a] I3M, Université Montpellier 2, Place Eugène Bataillon, 34090 Montpellier, France
[b] ALTADIS, Centre de recherche SCR, 4 rue André Dessaux, 45404 Fleury les Aubrais, France
[c] LISE CEREMADE, Université Paris IX Dauphine, Place de Lattre de Tassigny, 75016 Paris, France

ARTICLE INFO

Article history:
Received 17 August 2008
Received in revised form 7 March 2009
Accepted 11 March 2009
Available online 24 March 2009

Keywords:
Linear regression
Latent variables
Multi-block component regression model
PLS path modeling
PLS regression
Structural equation models
SEER

ABSTRACT

In this work, we consider chemical and physical variable groups describing a common set of observations (cigarettes). One of the groups, minor smoke compounds (*minSC*), is assumed to depend on the others (*minSC* predictors). PLS regression (PLSR) of *m* inSC on the set of all predictors appears not to lead to a satisfactory analytic model, because it does not take into account the expert's knowledge. PLS path modeling (PLSPM) does not use the multidimensional structure of predictor groups. Indeed, the expert needs to separate the influence of several pre-designed predictor groups on *minSC*, in order to see what dimensions this influence involves. To meet these needs, we consider a multi-group component-regression model, and propose a method to extract from each group several strong uncorrelated components that fit the model. Estimation is based on a global multiple covariance criterion, used in combination with an appropriate nesting approach. Compared to PLSR and PLSPM, the structural equation exploratory regression (SEER) we propose fully uses predictor group complementarity, both conceptually and statistically, to predict the dependent group.

1. Introduction

1.1. Context

Cigarette minor smoke compounds must comply with regulations. But chemical analysis of smoke is very difficult to perform and time-consuming compared to cigarette physical and chemical analysis. So, we aim at modeling minor smoke compounds directly from cigarette chemical and physical descriptors, as far as possible. There is still a possibility that cigarette chemical and physical descriptors be not sufficient to predict minor smoke characteristics, and should be complemented with variables capturing the effects of the smoking machine. Major smoke descriptors may then be used as proxy for such effects.

1.2. Data and problem

Our data consist in the following 4 numerical variable groups, describing 28 cigarettes, which make up a sample representative of the European market:

TChem group: 29 chemical tobacco variables.
CPhys group: 10 physical cigarette design variables.
MajSC group: 5 major smoke compounds.
minSC group: 14 minor smoke compounds.

Note that the data supplied by Altadis being confidential, variable labels are obtained by merely suffixing a number to the label of the variable group.

Tobacco chemists assume that some (unknown) aspects of *minSC* depend, to a certain (unknown) extent, on some (unknown) aspects of *TChem*, *CPhys* and *MajSC* (cf. Fig. 1).

Chemical tobacco variables and physical cigarette design ones must be separated because we are looking for chemical structures and physical structures, *respectively*: their explanatory/predictive powers should be distinguished as far as possible. So, predictors are to be partitioned into *conceptually homogeneous* groups which will be called *Thematic groups*, or *Themes*, whose effects onto the dependent group should be *separated*.

Every thematic group may contain *several* unknown important underlying dimensions. The relation network between themes can be referred to as *thematic model*. The thematic model we considered ex ante is illustrated in Fig. 2. Chemists want to know if tobacco and cigarette analysis (relatively cheap) can be sufficient to predict *minSC*, or if one still has to go through major smoke analysis (much

* Corresponding author. Tel.: +33 467 143578.
E-mail address: bry@math.univ-montp2.fr (X. Bry).

0003-2670/$ – see front matter © 2009 Published by Elsevier B.V.
doi:10.1016/j.aca.2009.03.013

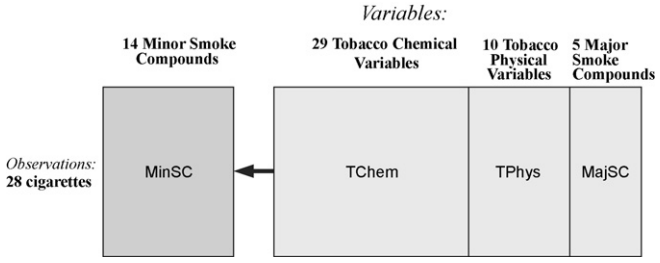

Fig. 1. Initial rough conceptual model.

more expensive) to predict *minSC*. Put it shortly, they want to know if any of the numbered links can be withdrawn from the thematic model shown in Fig. 2.

Chemists also wonder whether the conceptually appealing partitioning of predictors into *TChem*, *CPhys* and *MajSC* is really operational. Else, what alternative partitioning should they consider, both operationally more powerful and conceptually relevant?

In this paper, we only consider *minSC*'s model: we are not interested in the dotted arrows in Fig. 2. We want to extract the best possible strong dimensions in *TChem*, *C* Phys and *MajSC* to predict most of *minSC*.

1.3. Available methods: assets and limits

1.3.1. Gathering all predictor groups: the limits of classical PLS regression

The most obvious classical method available to investigate the multidimensional structure of both predictors and dependent variables with respect to the regression model seems to be PLSR. It may also seem fit to question the adequacy of the predictors' partition to the modeling purpose: if we mix up all predictors, and the predictive structures revealed by PLSR do not meet a given predictor partition, we tend to think that this partition should no longer be considered. But indeed, this is not the case: let us for example imagine a situation involving two predictor groups *A* and *B*, each being a tight variable bundle structured around one direction, the two directions being somewhat correlated, though not very strongly (cf. Fig. 3). From the conceptual point of view, such a situation may be regarded as ideal: each bundle can be assumed to measure a unidimensional concept with little error, and, as these bundles are weakly correlated, their effects on the dependent variables may be well separated. What happens with PLSR components is that, being constrained to be orthogonal, they will never adjust these correlated bundles correctly. Worse than that: supposing *A* and *B* have balanced effects on the dependent group, the first PLSR component is likely to stand in a position intermediate between the bundle directions. Then, because of orthogonality, so does the second (cf. Fig. 3). With more than two predictor groups, plane graphs using PLSR explanatory components may even capture no bundle direction at all.

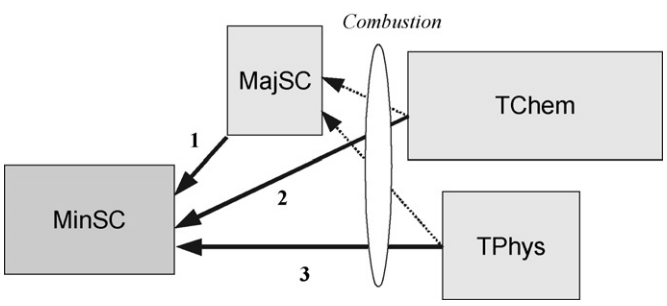

Fig. 2. Thematic model based on the chemists' a priori partition of predictors.

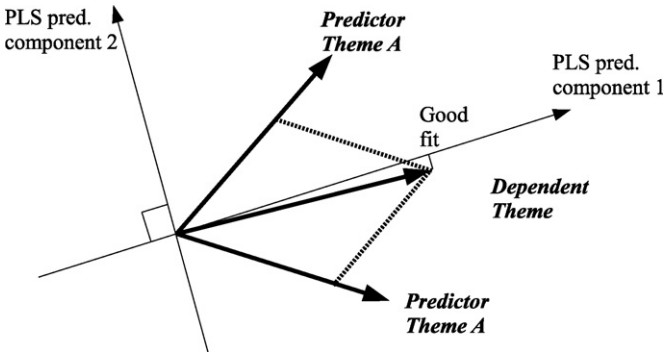

Fig. 3. How PLSR components may miss the thematic structures.

Besides, one should be aware that the higher the number of variables measuring a given aspect of observations, the heavier the weight of this aspect in the PLSR calculus of components. This is a major drawback when conceptually distinct aspects are mixed up, each being measured through a more or less arbitrary number of partially redundant variables, which happens in many cases. Indeed, the number of variables measuring an aspect often seems less related to the importance of this aspect than to the difficulty to measure it.

Finally, even when PLSR reveals predictive variable bundles, the conceptual mix up of predictors usually makes their interpretation more difficult.

In order to investigate the predictors' multidimensional structure and show that grouping predictors causes lack of explanatory power, PLSR was applied to our data, using *minSC* as dependent group and $X = \{TChem, TPhys, MajSC\}$ as sole predictor group. Detailed results are given in Section 4.

Next, we review some component regression methods taking into account a given partition of predictors. For each, we briefly point out its assets and limits.

1.3.2. Methods dealing with a predictor partition: state of the art

- *Principal component analysis* (PCA) performed separately on each thematic group does extract hierarchically ordered and uncorrelated principal dimensions in the theme, but regardless of the role they play in the thematic model.
- *PLS path modeling* (PLSPM) considers this role (see [1–3]). It basically assumes that each group is structured around a single latent variable (LV) and that LV's are linked together through a regression model. Each LV is to be estimated through a component. The PLSPM algorithm is fast and may deal with small samples. But it's estimation procedure is based on no global criterion. Besides, the fact that PLSPM extracts but one dimension per explanatory group prevents from graphing and exploring group structure. PLSPM was applied to our data, using *minSC* as dependent group and *TChem*, *TPhys* and *MajSC* as predictor groups. Results are given in Section 4.
- Jöreskog's *Structural equation model maximum of likelihood* (SEM-ML) [4], based on a global likelihood maximization involving latent variables, is theoretically better grounded, but requires heavy probabilistic assumptions, many observations, and sometimes leads to convergence problems. Just as PLSPM, SEM-ML extracts but one dimension in each thematic group.

  For a comparison of SEM-ML and PLS-PM, see [4].
- Such is not the case of the *multi-block PLS* (MBPLS) algorithm initially proposed by [5] and improved by [6–8]. It does not require probabilistic assumptions or many observations, and extracts several components per block. But on the other hand, for want of a global criterion to be optimized, this technique does not deal with partial relations appropriately. Besides, the algorithm is made