Description And Prediction From Multiblock Tables. Application To Epidemiological Data.

Bougeard S.[1], Chauvin C.[1], Hanafi M.[2] and Qannari E.M.[2]
[1] **Afssa**, Département d'épidémiologie animale - Les Croix, BP53, 22440 Ploufragan
[2] **Enitiaa-Inra**, Unité de Sensométrie et Chimiométrie - Rue de la Géraudière BP 82225, 44322 Nantes

ABSTRACT

Epidemiological data are usually organized in several tables. For example, the expression of an animal disease could be explained by data organized in four blocks related to feeding, hygiene, farming feature or treatments. All the variables of each table are measured on the same animals or farms. The first issue of the statistical treatment is to describe the multiblock tables and to sum up the complex links between the variables and between the data sets. The second issue is to predict the disease from the multiblock explanatory tables and to assess which variable and which block are good predictors of the disease. We purpose a new method, called Multiblock Latent Root Regression, where each data set is summarized with latent variables, linear combinations of variables derived from the data sets. These latent variables are computed by maximising a criterion based on covariance. They are directly used as predictors to explain the disease. It shares certain common characteristics with Partial Least Square regression (Wold et al., 1983) and Partial Least Square path modelling (Wold et al., 1996) as it derives latent variables to be used as predictors. The interest of the method is illustrated on the basis of a data set pertaining to epidemiology.

INTRODUCTION

Research in epidemiology is concerned with detecting, identifying and preventing animal diseases. Very often, the purpose of the study is to predict one or more variables related to animal health or the spread of an infection from variables related to the breeding environment, alimentary factors and farm management, amongst others. Disease is the result of interactions between the infectious agent, the affected animal, the environment and management factors. The explanatory variables (risk factors) may be directly observed on the animals or measured by means of questionnaires.

This paper deals with the description and the prediction of multiblock data. Very often, epidemiological data are usually organised in *(K+1)* blocks consisting of *K* explanatory blocks X_k and a *Y* data set to be predicted. All these measurements are made on the same animals or on the same farms. Latent root regression*, LRR,* (Webster et al., 1974) is an appropriate method to link a variable to a set of predictors. This method shares common characteristics with Partial Least Square *(PLS)* regression (Wold et al., 1983) as it derives latent variables to be used as predictors. Moreover, as in *PLS* regression, the variable to be predicted plays a central role in the determination of the latent variables. We discuss a new formulation of this method which emerges from a slight change in the optimisation criterion. It is simpler than the original method and can be easily extended to the multiblock setting.

METHOD: MULTIBLOCK LATENT ROOT REGRESSION

Resolution

The Multiblock Latent Root Regression, *M-LRR*, (Bougeard et al., Submitted) is adapted to the multiblock setting where we have *(K+1)* data sets: a data set *Y* to be predicted from *K* data sets X_k *(k=1, …, K).* The *Y* table contains *Q* variables and each table X_k contains P_k variables. All these variables are measured on the same individuals and supposed to be centred. In a first stage, we seek latent variables which are linear combinations of variables in the various data sets: $u^{(1)}=Yv^{(1)}$, $t_k^{(1)}=X_k w_k^{(1)}$ and $t^{(1)}=Xw^{(1)}$ where $X=[X_1|\ldots|X_K]$. These latent variables are sought in such a way so as to maximize the following criterion based on the covariance between each partial latent variable and the global latent variable:

$$\operatorname{cov}^2(u^{(1)},t^{(1)})+\sum_{k=1}^{K}\operatorname{cov}^2(t_k^{(1)},t^{(1)}) \text{ with } t^{(1)}=Xw^{(1)}, u^{(1)}=Yv^{(1)}, t_k^{(1)}=X_k w_k^{(1)} \text{ and } \left\|w^{(1)}\right\|=\left\|v^{(1)}\right\|=\left\|w_k^{(1)}\right\|=1$$

Using Cauchy-Schwarz inequality, we prove that the optimal vector of loadings $w^{(1)}$ is given by the eigenvector of the matrix $(X'YY'X+X'XX'X)$ where $X=[X_1|\ldots|X_K]$. Then the components $t^{(1)}$, $u^{(1)}$, $t_k^{(1)}$, and the vectors of loadings $v^{(1)}$ and $w_k^{(1)}$, are directly derived from the eigenvector $w^{(1)}$: $t^{(1)}=Xw^{(1)}$, $v^{(1)}=Y't^{(1)}/||Y't^{(1)}||$, $u^{(1)}=Yv^{(1)}$, $w_k^{(1)}=X_k't^{(1)}/||X_k't^{(1)}||$ and $t_k^{(1)}=X_k w_k^{(1)}$. In order to determine the second order solution, we consider

the orthogonal projection $X_k^{(1)}$ (respectively $Y^{(1)}$) of X_k (respectively Y) onto the subspace spanned by the first global component $t^{(1)}$. $X^{(1)}$ is defined as $X^{(1)}=[X_1^{(1)}|\ldots|X_K^{(1)}]$. The optimal vector of loadings $w^{(2)}$ is given by the eigenvector of $(X^{(1)\prime}Y^{(1)}Y^{(1)\prime}X^{(1)}+X^{(1)\prime}X^{(1)}X^{(1)\prime}X^{(1)})$ associated with the largest eigenvalue. The partial latent variables $u^{(2)}$ and $t_k^{(2)}$ are computed in a similar way as previously. This process can be reiterated in order to find higher order solutions.

Prediction

The prediction of the Y variables is based on the global latent variables $(t^{(1)}, \ldots, t^{(h)})$. These latent variables being orthogonal by construction, the Y table is split up into orthogonal components: $Y=t^{(1)}c^{(1)\prime}+\ldots+t^{(h)}c^{(h)\prime}+Y^{(h)}$. Moreover, the global latent variables can be expressed as linear combinations of X: $t^{(1)}=Xw^{*(1)},\ldots, t^{(h)}=Xw^{*(h)}$. The loadings $w^{*(h)}$ and $c^{(h)}$ are defined as in *PLS* regression (Tenenhaus, 1998). This leads to the model: $Y=X(w^{*(1)}c^{(1)\prime}+\ldots+w^{*(h)}c^{(h)\prime})+Y^{(h)}$. The choice of the optimal number of latent variables to be introduced in the model is based on a cross-validation procedure (Stone, 1974).

APPLICATION

Multiblock Data And Objectives

The data set consists in the measurements on *659* turkey flocks. The variables are organised in *3* tables. The Y table, which contains *2* variables, concerns the farmer loss in terms of mortality and condemnation at slaughterhouse. The X explanatory table is organised in *2* blocks: X_1 and X_2 (description in table 1). Table X_1 contains *14* variables pertaining to the farming features. Table X_2 contains *5* variables which refer to technical and economical results of the flocks. Indicator variables are considered for the categorical variables. Because all variables have different scales, they are centred and scaled. Moreover, each data set is accommodated by an isotropic scaling factor. The two aims of the statistical analysis are to determine which variables can differentiate the turkey flocks and to predict the Y table from X_1 and X_2.

Description Of Multiblock Data Sets

The first global component $t^{(1)}$, which represents *28.7%* of the total variance in both X and Y variables, is highly related to data table X_2 as it explains *55.2%* of the total variance in this data set. It also explains *19.2%* and *20.9%* of the total variance in X_1 and Y respectively. The second global latent variable $t^{(2)}$ explains up to *44.4%* of the total variance in Y. It turns out that the first global latent variable $t^{(1)}$ reflects a common structure of both the data sets X_1 and X_2 whereas the second global latent variable is mainly related to X_2. It also reflects the importance of data set X_2 in the prediction of Y. Figure 1 shows how the variables and the latent variables obtained at the first and second order are related to each others. It turns out that the carcasses condemnation (*CONDEMN*) and the variables *DENSI*, *MOYVET* and *VET* in data table X_1 and *KGM2*, *DWG*, *RI*, and *TCI* in data table X_2 are best explained by $t^{(1)}$. The second variable in Y (*MORT*) and the variables *ECI* and *TCI* from X_2 are best reflected in the second latent variable $t^{(2)}$.

Prediction From Multiblock Data Sets

As discussed above, a prediction model can be set up by regressing the Y variables on the basis of the global latent variables. A cross-validation procedure was performed in order to assess the appropriate number of latent variables to be introduced in the model. This led to the choice of *5* latent variables. Table 1 shows the regression coefficients associated with the variables in Y, loss in terms of mortality (*MORT*) and condemnation at slaughtering (*CONDEMN*). It turns out that the X variables which most influence Y are in particular the serious health problem during farming (*SANPB*), the economical consumption index (*ECI*) and the technical consumption index (*TCI*). A hypothesis that emerges from these findings is that the condemnation at slaughtering is linked to health problems onset and more generally to the technical performance of the breeders. The mortality during farming (*MORT*) is mainly predicted with three variables from X_2: the economical consumption index (*ECI*), the technical consumption index (*TCI*) and the result index (*RI*). This means in particular that the mortality is reflected by some economical performance indices.

CONCLUSION

The Multiblock-Latent Root Regression might constitute solutions to overcome practical problems of complex epidemiological data. Firstly, this technique is little sensitive to the quality of data, for example sample size and multi-collinearity in the explanatory data set. Then we can use more than one variable to predict and it is more and more frequent to describe complex diseases. This table to be predicted plays an active role in the analysis. This ensures that major principal axes will be related to these variables and then that the factorial representation is oriented towards the disease explanation. Moreover, modelling and factorial representation are performed at the same time as in *PLS* regression. Finally, we can take into account that the explanatory tables are organized in several meaningful blocks and measure the block weights in the disease explanation.

The rational behind Multiblock-LRR is easy to understand because there is a global criterion to maximise and, moreover, the solutions are derived from an eigenanalysis of a matrix. The advantage of the method lies in the introduction of global latent variables on the one hand and partial latent variables on the other hand. These latent variables highlight the relationships among the various data sets. In particular, the introduction of the global latent variables is very useful from a practical point of view as illustrated in the case study, but also from a conceptual standpoint because it provides a straightforward way to derive orthogonal latent variables. Needless to mention that the approach can easily be extended to the case where the *Y* table is organised into multiblock data sets.

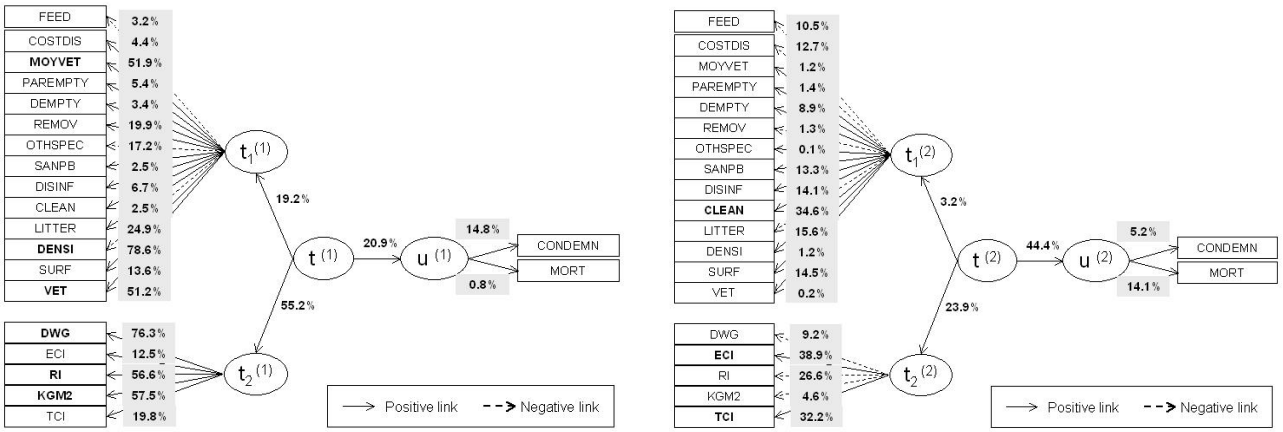

Figure 1: Scheme highlighting the relationships between the variables and their associated partial latent variables, in terms of percentage of explained variance, for the two first dimensions.

Table 1: Regression coefficients of *Y* on *X=[X₁|X₂]* using *5* global latent variables retained.

Block	Variable	Variable description	CONDEMN	MORT
	FEED	Eat and bone meal-free feeding (1=yes, 0=no)	*-0.27*	*-0.12*
	COSTDIS	Disinfection costs	*-0.33*	*-0.02*
	MOYVET	Average veterinary costs for the last three flocks	***0.39***	*0.05*
	PAREMPTY	Partial emptying (1=yes, 0=no)	*0.25*	*-0.06*
	DEMPTY	Duration of the empty period before chick arrival	*0.11*	*0.11*
	REMOV	Number of removal to slaughterhouse per flock	*0.05*	*-0.06*
X₁	OTHSPEC	Last flock with the same species (1=yes, 0=no)	*0.02*	*-0.05*
	SANPB	Serious health problem during farming (1=yes, 0=no)	***0.66***	*0.18*
	DISINF	Disinfection labour (1=skilled labour, 0=yourself)	*0.03*	*0.23*
	CLEAN	Cleaning labour (1=skilled labour, 0=yourself)	*0.06*	*-0.24*
	LITTER	Quantity of litter used for the flock	*-0.18*	*-0.21*
	DENSI	Chick density at the beginning of farming	*-0.13*	*-0.06*
	SURF	Surface area on which the flock is farmed	*0.24*	*0.17*
	VET	Total amount of veterinary costs for the flock	***0.34***	*0.01*
	DWG	Daily weight gain	*0.11*	*-0.18*
	ECI	Economical consumption index	*0.20*	***0.55***
X₂	RI	Result index	*-0.10*	***-0.37***
	KGM2	Total flock weight slaughtered related to the surface area	*-0.02*	*-0.24*
	TCI	Technical consumption index	*0.19*	***0.45***

REFERENCES

Bougeard, S., Hanafi, M. and Qannari, E. M. (Submitted) Multiblock latent root regression. Application to epidemiological data, *Computational statistics and data analysis*.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society,* 36**,** 111-147.

Tenenhaus, M. (1998) *La régression PLS. Théorie et pratique,* Technip, Paris.

Webster, T., Gunst, R. F. and Mason, R. L. (1974) Latent root regression analysis, *Technometrics,* 16**,** 513-522.

Wold, S., Kettaneh, N. and Tjessem, K. (1996) Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection, *Journal of chemometrics,* 10**,** 463-482.

Wold, S., Martens, H. and Wold, H. (1983) The multivariate calibration problem in chemistry solved by the PLS method, *Proceedings of the Conference on Matrix Pencils***,** 286-293.