Noname manuscript No.
(will be inserted by the editor)

Partitioning predictors in multivariate regression models

Francesca Martella · Donatella Vicari · Maurizio Vichi

Received: date / Accepted: date

Abstract A Multivariate Regression Model Based on the Optimal Partition of Predictors (MRBOP) useful in applications in the presence of strongly correlated predictors is presented. Such classes of predictors are synthesized by latent factors, which are obtained through an appropriate linear combination of the original variables and are forced to be weakly correlated. Specifically, the proposed model assumes that the latent factors are determined by subsets of predictors characterizing only one latent factor. MRBOP is formalized in a least squares framework optimizing a penalized quadratic objective function through an alternating least-squares (ALS) algorithm. The performance of the methodology is evaluated on simulated and real data sets.

Keywords Penalized regression model · Partition of variables · Least squares estimation · Class-correlated variables · Latent factors

1 Introduction

Applications where several dependent variables (responses) have to be predicted using a large number of variables have been considered in various disciplines such as bioinformatics, brain imaging, data mining, genomics and economics (Frank and Friedman, 1993; Waldro et al., 2011). The standard model that accommodates this issue is the ordinary multivariate regression model, see chapter 15 in Krzanowski (2000). Fitting the multivariate regression model to observed data requires estimation of the unknown regression coefficients and the dispersion matrix of the error terms. This can be done by maximum likelihood if normality is assumed for the error matrix or by least-squares if no distributional assumptions are made. For standard application of the inferential theory, it is necessary for the sample size to be greater than the number of predictors plus the number of responses and for the predictor matrix to be of full rank. However, when the number of predictors is large, two important statistical problems, which are related to each other, may commonly arise:

1. difficulty in interpretation of the (many) regression coefficients;
2. presence of strongly correlated predictors.

In the latter case, the regression approach may still be able to determine the ordinary least squares (OLS) estimators, but it may not be able to distinguish the effect of each predictor on the responses (multicollinearity). In the literature many proposals and strategies for dealing with such problems have been proposed, which can be classified into three categories: standard variable selection methods, see as an example Hocking (1976), penalized (or shrinkage) techniques, also known as bias estimation, see Tibshirani (1996); Hoerl and Kennard (1970); Frank and Friedman (1993); Zou and Hastie (2005); Tutz and Ulbricht (2009); Witten and Tibshirani (2009); Yuan and Lin (2006), and dimensionality reduction methods (DRMs).

In particular, we focus on DRM regression where a small set of linear combinations of the original vari-

Francesca Martella
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale Aldo Moro, 5 - 00185 Rome (Italy)
Tel.: +39-06-49910464
E-mail: francesca.martella@uniroma1.it

Donatella Vicari
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale Aldo Moro, 5 - 00185 Rome (Italy)

Maurizio Vichi
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale Aldo Moro, 5 - 00185 Rome (Italy)

ables are built and then used as input to the regression model. DRMs differ in how the linear combinations are built. For example, principal component regression (PCR) (Jolliffe, 1982) performs a PCA on the explanatory variables and uses the principal components as latent predictors in the regression model. It has been demonstrated that this strategy does not guarantee the principal components, which optimally "explain" the predictors, will be relevant for the prediction of the responses. Canonical correlation regression (CCR) (Hotelling, 1935) is similar to PCR, but whereas PCR finds the directions of maximal variance of each predictor-space, CCR identifies directions of maximal correlation of both predictor and response spaces. Partial least squares regression (PLSR) (Wold, 1966) represents a form of CCR where the criterion of maximal correlation is balanced by requiring the model to explain as much variance as possible in both predictor and response spaces. Stone and Brooks (1990) proposed continuum regression as a unified regression technique embracing OLS, PLSR and PCR. On the other hand, Reduced Rank Regression (RRR) (Anderson, 1951; Izenman, 1975) minimizes the sum of the squared residuals subject to a reduced rank condition. It can be seen as a regression model with a coefficient matrix of reduced rank. Moreover, it can be shown that the RRR latent predictors are the same as the ones from the Redundance Analysis (RA), see Van Den Wollenberg (1977). Yuan et al. (2007) proposed a general formulation for dimensionality reduction and coefficient estimation in multivariate linear regression, which includes many existing DRMs as specific cases. Bougeard et al. (2007) proposed a new formulation to the multiblock setting of latent root regression applied to epidemiological data and Bougeard et al. (2008) investigated a continuum approach between MR and PLS. However, dimensionality reduction methods, such as PCA and Factor Analysis, may generally suffer from a lack of interpretability of the resulting linear combinations. Rotation methods are often used to overcome such a problem. In the regression context, we propose here to simplify the interpretation by partitioning the predictors into classes of correlated variables synthesized by weakly correlated factors that best predict the responses in the least-squares sense. This turns out to be a relevant gain in the interpretation of the regression analysis, which can be nicely displayed by a path diagram identifying the underlying relations between predictors, latent factors and responses. The model should be used when the researcher does not have a priori hypotheses about the association patterns present between the manifest variables.

In this paper a multivariate regression model based on the optimal partition of predictors (MRBOP) is proposed which optimally predict the responses in the least-squares sense. In fact, the assumption underlying the model is the existence of weakly correlated groups of predictors but the number and the composition of such possible blocks need to be determined. In the framework of DRMs, the new methodology determines latent factors as linear combinations of such subsets of predictors by performing simultaneously the clustering of the predictors and the estimation of the regression coefficients of the derived latent factors. Specifically, the model proposed aims at defining classes of correlated predictors, which lead to the construction of weakly correlated latent factors. This could be particularly useful in high dimensional regression studies where strongly correlated variables might represent an unknown underlying latent dimension. In these cases, methods that are commonly employed as cures for collinearity - in particular, variable selection - can be inadequate because important grouping (i.e. latent dimension) information may be lost. Moreover, in the case of perfect linear relationship among predictors, we can here algebraically derive the regression coefficient estimators contrarily to the OLS framework. Finally, an important advantage of MRBOP is the interpretability of each latent factor representing only one subset of well-characterized predictors. In fact, predictors are not allowed to influence more than one factor as frequently happens in dimensionality reduction methods.

The model is formalized in a least squares estimation framework optimizing a penalized quadratic objective function. The paper is organized as follows. Section 2 describes the general DRMs and discusses the possible specifications, while Section 3 introduces the model and an alternating least-squares algorithm to estimate the model parameters. An illustrative example is presented in Section 3.2.1. In Sections 4 and 5, the results obtained on simulated and real data sets are discussed. The last Section is devoted to concluding remarks.

2 Dimensionality reduction methods in multivariate regression

Let $\mathbf{X} = [x_{ij}]$ be a $(I \times J)$ matrix, where x_{ij} represents the value of the j-th predictor observed on the i-th subject and $\mathbf{Y} = [y_{im}]$ be a $(I \times M)$ matrix, where y_{im} is the value of the m-th response observed on the i-th subject. Without loss of generality, after a location and scale transformation, we can assume that all the variables are centered and standardized. As mentioned in the Introduction, several methods have been proposed to overcome problems connected with the presence of a

relatively large number of predictors. In particular, an attractive class of methods is represented by the DRMs where the responses are regressed against a small number $Q \leq J$ of latent factors obtained as linear combinations of the original predictors. These methods can be expressed as follows:

$$\mathbf{Y} = \mathbf{ZC} + \mathbf{E} \tag{1}$$

where $\mathbf{Z} = \mathbf{X}\tilde{\mathbf{V}}$ represents the $(I \times Q)$ matrix of the latent predictors, $\tilde{\mathbf{V}}$ being the $(J \times Q)$ unknown factor loading matrix, $\mathbf{C}$ is the $(Q \times M)$ regression coefficient matrix and $\mathbf{E}$ is the $(I \times M)$ matrix of unobserved random disturbances. As usual in the least-squares context, we do not impose any distributional assumption on $\mathbf{E}$. Clearly, equation (1) can be re-written as

$$\mathbf{Y} = \mathbf{X}\tilde{\mathbf{V}}\mathbf{C} + \mathbf{E} = \mathbf{XB} + \mathbf{E} \tag{2}$$

where $\mathbf{B} = \tilde{\mathbf{V}}\mathbf{C}$ is the $(J \times M)$ matrix of the regression coefficients of the Q latent factors in the original space. Unfortunately, the decomposition $\mathbf{B} = \tilde{\mathbf{V}}\mathbf{C}$ is not unique, since for every invertible $(Q \times Q)$ matrix $\mathbf{F}$, we have

$$\mathbf{B} = \tilde{\mathbf{V}}\mathbf{C} = \tilde{\mathbf{V}}\mathbf{F}\mathbf{F}^{-1}\mathbf{C} = (\tilde{\mathbf{V}}\mathbf{F})(\mathbf{C}'\mathbf{F}'^{-1})'. \tag{3}$$

In the next section, we discuss different proposals to solve such an identifiability issue.

2.1 Possible specifications for DRMs

A first raw specification of the model is represented by PCR, which is based on a two-step procedure. It first selects the principal component matrix $\mathbf{Z} = \mathbf{X}\tilde{\mathbf{V}}$, and then uses such latent variables as predictors of the regression model predicting $\mathbf{Y}$. Hence, the columns of $\tilde{\mathbf{V}}$ are represented by the first Q eigenvectors normalized to length one associated with the largest eigenvalues of $\mathbf{X}'\mathbf{X}$. However, this approach does not guarantee that the principal components, which optimally explain $\mathbf{X}$, will be relevant for the prediction of $\mathbf{Y}$.
A more refined solution to solve the problem, which is at the base of many DMRs, is obtained by constraining $\tilde{\mathbf{V}}'\mathbf{X}'\mathbf{X}\tilde{\mathbf{V}} = \mathbf{I}_Q$ as in PCR, and simultaneously selecting the latent factors which best predict the responses through the following least-squares problem

$$\| \mathbf{Y} - \mathbf{ZC} \|^2 . \tag{4}$$

It has to be noticed that given matrix $\mathbf{Z}$, the regression coefficient matrix $\mathbf{C}$ turns out to be the OLS solution of (1), i.e. $\hat{\mathbf{C}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$. Therefore, to estimate the regression coefficients $\mathbf{B} = \tilde{\mathbf{V}}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$ it is sufficient to estimate $\mathbf{Z}$. The DMRs differ in the way the latent factors, and therefore $\tilde{\mathbf{V}}$, are obtained.
A popular example of DMRs methods is represented by RRR (RA), where the latent predictors are constrained to be orthogonal to each other and have unit length. In particular, the columns of $\tilde{\mathbf{V}}$ correspond to a set of redundant latent variables and are given by the first Q eigenvectors associated to the largest eigenvalues of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}$.
Another model which is closely related to RRR is CCR, which provides the generalized least-squares solutions to the RRR model as latent predictors. In details, the columns of $\tilde{\mathbf{V}}$ are given by the first Q coefficients of the canonical correlation variables in the predictor space, that is by the eigenvectors corresponding to the first Q largest eigenvalues of the matrix $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{X}$.
A mixture of RRR and PCR (Abraham and Merola, 2005) is represented by PLSR. In the literature different versions of PLSR exist, see Rosipal and Krmer (2006) and Abdi (2010). However, PLS produces similar results to its variant called SIMPLS (De Jong, 1993), and, for one response, the results are even identical. In particular, SIMPLS maximizes $(\tilde{\mathbf{v}}'\mathbf{X}'\mathbf{y})^2$ under the constraint that the dimensions $\mathbf{X}\tilde{\mathbf{v}}$ are orthogonal to each other, and that $\tilde{\mathbf{v}}'\tilde{\mathbf{v}}$=1. As it turns out, the loading associated with the first latent factor ($\tilde{\mathbf{v}}^{(1)}$) is determined as the first eigenvector of $\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}$. Subsequent latent variables may be obtained by iterative deflations. Moreover, it could be demonstrated that for spherically distributed input data, PLS produces the same result as RRR.

3 MRBOP model

One of the most important problems related to the DRMs, but more in general to all reduction methods based on latent variables, is the interpretation of the latent factors. A standard way to proceed is to look at the loadings $\tilde{\mathbf{v}}_{jq}$ (each q-th latent factor is characterized by the original predictors corresponding to the highest absolute values of $\tilde{\mathbf{v}}_{jq}$) or to look at the correlation coefficients between latent and original variables. However, those procedures are heuristic and not always applicable because in practical situations the original predictors may have more than one high loading (or correlation) value for several latent predictors, especially when a large number of explanatory variables are considered and relatively high correlations are present in the data.
In this respect, the starting point of our proposal is that $\tilde{\mathbf{V}}$ is assumed to be a column-orthonormal matrix (i.e. $\tilde{\mathbf{V}}'\tilde{\mathbf{V}} = \mathbf{I}_Q$) having a particular structure with only one non-null element per row. In particular, $\tilde{\mathbf{V}}$ is parameterized as follows:

$$\tilde{\mathbf{V}} = \mathbf{WV} \tag{5}$$

where $\mathbf{W}$ is a $(J \times J)$ diagonal matrix which gives weights to the J predictors and $\mathbf{V}$ is a $(J \times Q)$ binary and row stochastic matrix defining a partition of the predictors in Q non-empty classes.
By including (5) into model (2), the proposed MRBOP model is specified by

$$\mathbf{Y} = \mathbf{XWVC} + \mathbf{E}, \tag{6}$$

subject to

$v_{jq} \in \{0, 1\}$, $j = 1, ..., J$; $q = 1, ..., Q$;
$\sum_{q=1}^{Q} v_{jq} = 1$, $j = 1, ..., J$;
$\mathbf{W}$ is a diagonal weight matrix, such that $(\mathbf{WV})'\mathbf{WV} = \mathbf{I}_Q$.

Note that the factor loading matrix $\tilde{\mathbf{V}}$ can be rotated without affecting the model, provided that the regression coefficient matrix is counter-rotated by the inverse transformation.
Model (6) implies that the Q latent factors are easily interpretable in terms of the J original variables because each of the latent factors is a linear combination of only one subset of the correlated predictors.
Furthermore, it is worthwhile to note that matrix $\mathbf{V}$, being binary and row stochastic, univocally defines a partition of the predictors that should be identified: i) to best predict the responses; ii) to include correlated predictors within classes and weakly correlated predictors between classes.
In order to specify mathematically property ii), model (6) needs some requirements to force the solution towards such a direction. In particular, given a partition identified by a matrix $\mathbf{V}$, let the correlation matrix of $\mathbf{X}$

$$\mathbf{R} = \frac{1}{I}\mathbf{X}'\mathbf{X}$$

be decomposed into matrix $\mathbf{R_W}$, whose non-null entries are the correlations between predictors within classes,

$$\mathbf{R_W} = \sum_{q=1}^{Q} diag(\mathbf{v}_q)\mathbf{R}diag(\mathbf{v}_q)$$

and matrix $\mathbf{R_B}$, where the non-null entries are the correlations between predictors belonging to different classes,

$$\mathbf{R_B} = \sum_{q,p=1,q\neq p}^{Q} diag(\mathbf{v}_q)\mathbf{R}diag(\mathbf{v}_p).$$

Thus, resulting in the formula

$$\mathbf{R} = \mathbf{R_W} + \mathbf{R_B}.$$

The trace of the squared covariance matrix is the scalar-valued variance (denoted by VAV) and the trace of the product of two covariance matrices is the scalar-valued covariance, denoted by $COVV$ (Escoufier, 1973). Therefore, the following three measures

$$\| \mathbf{R} \|^2 = tr(\mathbf{RR}) = \frac{1}{I^2} \| \mathbf{X}'\mathbf{X} \|^2 = VAV(\mathbf{X}); \tag{7}$$

$$\sum_{q=1}^{Q} \| diag(\mathbf{v}_q)\mathbf{R}diag(\mathbf{v}_q) \|^2 = tr(\mathbf{R_W}\mathbf{R_W}) = \tag{8}$$
$$= \sum_{q=1}^{Q} VAV(\mathbf{X}diag(\mathbf{v}_q));$$

$$\sum_{q,p=1,q\neq p}^{Q} \| diag(\mathbf{v}_q)\mathbf{R}diag(\mathbf{v}_p) \|^2 = tr(\mathbf{R_B}\mathbf{R_B}) = \tag{9}$$
$$= \sum_{q,p=1,q\neq p}^{Q} COVV(\mathbf{X}diag(\mathbf{v}_q), \mathbf{X}diag(\mathbf{v}_p));$$

are multivariate measures of the variance of all the predictors, $\mathbf{X}$, of the predictors within the same class, $\mathbf{X}diag(\mathbf{v}_q)$, and, of the covariance of the predictors belonging to different classes, $\mathbf{X}diag(\mathbf{v}_q)$ and $\mathbf{X}diag(\mathbf{v}_p)$, respectively.
Note that the term in the sum in (8)

$$\mathbf{R}_{\mathbf{W}_q} = diag(\mathbf{v}_q)\mathbf{R}diag(\mathbf{v}_q)$$

is just the covariance matrix of the predictors within the q-th class. Since (8) increases as the class sizes increase, we may want to weigh each class by taking into account its size:

$$\frac{1}{n_q}\mathbf{R}_{\mathbf{W}_q} = \frac{1}{n_q}diag(\mathbf{v}_q)\mathbf{R}diag(\mathbf{v}_q)$$

where n_q denotes the size of the q-th cluster. Hence, writing $\mathbf{N} = diag(\mathbf{VV}')^{-1}\mathbf{1}_J$, the weighted versions of (7), (8) and (9) are

$$\| \mathbf{RN} \|^2 = tr(\mathbf{NRRN}); \tag{10}$$

$$\sum_{q=1}^{Q} \| diag(\mathbf{v}_q)\mathbf{R}diag(\mathbf{v}_q)\mathbf{N} \|^2 = \tag{11}$$
$$= tr(\mathbf{N}\mathbf{R_W}\mathbf{R_W}\mathbf{N});$$

$$\sum_{q,p=1,q\neq p}^{Q} \| diag(\mathbf{v}_q)\mathbf{R}diag(\mathbf{v}_p)\mathbf{N} \|^2 = \tag{12}$$
$$= tr(\mathbf{N}\mathbf{R_B}\mathbf{R_B}\mathbf{N}).$$

Therefore, to achieve our goal of partitioning the predictors into classes formed by strongly correlated variables belonging to the same class or, equivalently, formed by

weakly correlated variables in different classes, we specify the least-squares estimation of model (6) as the solution of the following constrained quadratic problem:

$$\min_{\mathbf{W},\mathbf{V},\mathbf{C}} F(\cdot) = \min_{\mathbf{W},\mathbf{V},\mathbf{C}} \Big[\frac{\| \mathbf{Y} - \mathbf{XWVC} \|^2}{\| \mathbf{Y} \|^2} \Big] \tag{13}$$

subject to constraints

a) $v_{jq} \in \{0,1\}$, $j = 1, ..., J$; $q = 1, ..., Q$;
b) $\sum_{q=1}^{Q} v_{jq} = 1$, $j = 1, ..., J$;
c) $\mathbf{W}$ is a diagonal weight matrix, such that

$$(\mathbf{WV})'\mathbf{WV} = \mathbf{I}_Q;$$

d) $P = \frac{\sum_{q,p=1,q\neq p}^{Q} \| diag(\mathbf{v}_q)\mathbf{R} diag(\mathbf{v}_p)\mathbf{N} \|^2}{\|\mathbf{RN}\|^2} =$
$= \frac{tr(\mathbf{NR_B R_B N})}{tr(\mathbf{NRRN})} \leq S$

where S is a specified parameter. There is a one-to one correspondence between S in d) and parameter λ in the following penalized function:

$$\min_{\mathbf{W},\mathbf{V},\mathbf{C}} F(\cdot) = \min_{\mathbf{W},\mathbf{V},\mathbf{C}} \Big[\frac{\| \mathbf{Y} - \mathbf{XWVC} \|^2}{\| \mathbf{Y} \|^2} + \lambda P \Big] \tag{14}$$

where the first term represents the non-penalized least-squares problem normalized in [0,1] and λ is the positive penalty parameter. Note that, the higher value of λ is, the stronger the penalty is. The idea behind the penalty function d) is to force the non-penalized MRBOP function to identify classes formed by strongly correlated predictors or equivalently, classes having weak *between* correlations. The penalty function P approaches 0 when correlations between predictors in different classes approach 0; it assumes value 1 when correlations between predictors in the same group are null, while a value $P = 0.5$ means that the within correlations in matrix ($\mathbf{R_W}$) and the between correlations ($\mathbf{R_B}$) contribute equally to $\mathbf{R}$.
Given λ and Q, the minimization of the penalized problem (14) can be solved by using an alternating least-squares (ALS) algorithm, which iterates three main steps:

0. Initialization of $\mathbf{V}$ and $\mathbf{W}$
1. Updating $\mathbf{V}$ (Variable allocation step)
2. Updating $\mathbf{W}$ (Weighting step)
3. Updating $\mathbf{C}$ (Regression step)

3.1 Alternating least-squares (ALS) algorithm

– **Initialization**
Choose starting values for $\mathbf{V}$ and $\mathbf{W}$ randomly or in a rational way.

– **Updating V**
As far as the allocation step is concerned, the minimization of the penalized function (14) is achieved by solving an assignment problem with respect to $\mathbf{V}$, given $\hat{\mathbf{W}}$ and $\hat{\mathbf{C}}$, which is sequentially solved for the different rows of $\mathbf{V}$ by taking:

$$\hat{v}_{jq} = \begin{cases} 1 \text{ if } \quad F(\cdot, v_{jq} = 1) = \min_h F(\cdot, v_{jh} = 1) \\ 0 \text{ otherwise} \end{cases} \tag{15}$$

$j = 1, ..., J$, $q, h = 1, ..., Q$ and $q \neq h$. When $\hat{\mathbf{V}}$ is updated, a check to prevent from having possible empty classes is carried out.

– **Updating W**
Concerning the estimation of the diagonal weight matrix $\mathbf{W}$, given $\hat{\mathbf{V}}$ and $\hat{\mathbf{C}}$, different cases can be considered depending on possible restrictions on the entries of $\mathbf{W}$ due to parsimony requirements. In the simplest case the predictors in the same class q of size n_q have the same weights $\hat{w}_{jj}^q = \frac{1}{\sqrt{n_q}}$, $q = 1, ..., Q$ and $j = 1, ..., J$. A second case could be that predictors in the same class have same weights but possibly different signs, as for example under the constraints $w_{jj}^q = \{\frac{-1}{\sqrt{n_q}}, \frac{+1}{\sqrt{n_q}}\}$ ($j = 1, ..., J$ and $q = 1, ..., Q$). The estimation can be performed by sequentially assigning $\hat{w}_{jj}^q = \{\frac{1}{\sqrt{n_q}}$, if $F(\cdot, w_{jj}^q = \{\frac{1}{\sqrt{n_q}}) = \min\{F(\cdot, w_{ll}^q) : l = 1, ..., J\}$ and $\hat{w}_{jj}^q = \{\frac{-1}{\sqrt{n_q}}$ otherwise ($q = 1, ..., Q$ and $j = 1, ..., J$).
For the unconstrained weight matrix, where the predictors are free (in strength and sign) to differentially weigh in determining the new latent variables, the estimation of the diagonal values of $\mathbf{W}$, given $\hat{\mathbf{V}}$ and $\hat{\mathbf{C}}$ is done by rewriting the model

$$F(\cdot) \propto \| \mathbf{Y} - \mathbf{XWVC} \|^2 = \tag{16}$$
$$= \| \mathbf{Y} - \sum_{j=1}^{J} \mathbf{x}_j w_{jj} \tilde{\mathbf{c}}'_j \|^2 =$$
$$= \| \text{vec}(\mathbf{Y}) - \sum_{j=1}^{J} (\tilde{\mathbf{c}}_j \otimes \mathbf{x}_j) w_{jj} \|^2$$

where $\mathbf{x}_j$ is the j-th column vector of $\mathbf{X}$, $\tilde{\mathbf{c}}_j$ is the $(M \times 1)$ column vector representing the j-th row of $\mathbf{VC}$, and vec$(\cdot)$ and $\otimes$ denote the column vectorization of a matrix and the Kronecker product, respectively. The minimization of (14) is obtained to find the optimal diagonal entries w_{jj} ($j = 1, ..., J$) by solving an ordinary regression problem

$$F(\mathbf{w}) = \| \text{vec}(\mathbf{Y}) - \mathbf{Aw} \|^2$$

where $\mathbf{A}$ is the $(IM \times J)$ matrix having the Kronecker products of the corresponding columns of $\mathbf{VC}$ and $\mathbf{X}$ (i.e. the j-th column of $\mathbf{A}$ is $\tilde{\mathbf{c}}_j \otimes \mathbf{x}_j$)

as columns and $\mathbf{w}$ is the J-dimensional vector containing the elements w_{jj}. Clearly, the minimization problem is equivalent because the diagonal elements of $\mathbf{W}$ are the very elements of $\mathbf{w}$ and the optimal $\hat{\mathbf{w}}$ can be found as

$$\hat{\mathbf{w}} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\text{vec}(\mathbf{Y}),$$

whence the optimal $\hat{\mathbf{W}}$ is determinated by simply setting $\hat{\mathbf{w}}$ as its diagonal. In order to ensure column orthonormality of $\hat{\tilde{\mathbf{V}}} = \hat{\mathbf{W}}\hat{\mathbf{V}}$, the diagonal entries of $\hat{\mathbf{W}}$ need to be class-normalized: $\hat{w}^q_{jj} \longleftarrow \frac{\hat{w}^q_{jj}}{\sum_{l=1}^{J}(\hat{w}^q_{ll})^2}$ $(q = 1, ..., Q,; \; j = 1, ..., J)$ so that the constraint (13c) is fulfilled.

– **Updating C**
Finally, the estimation of the regression coefficient matrix $\mathbf{C}$, given $\hat{\mathbf{W}}$ and $\hat{\mathbf{V}}$, is performed easily because the problem turns into an ordinary unconstrained regression framework

$$\hat{\mathbf{C}} = (\hat{\mathbf{V}}'\hat{\mathbf{W}}\mathbf{X}'\mathbf{X}\hat{\mathbf{W}}\hat{\mathbf{V}})^{-1}\hat{\mathbf{V}}'\hat{\mathbf{W}}\mathbf{X}'\mathbf{Y}. \quad (17)$$

Given Q and λ, the three steps are alternated repeatedly until convergence, obtained with a sequence of values of the function $F(\cdot)$, which is bounded from below. To avoid the well known sensitivity of the ALS algorithms to the choice of starting values and to increase the chance of finding the global minimum, the algorithm should be run several times starting from different initial (random or rational) estimates of $\mathbf{V}$ and retaining the best solution, i.e. the one minimizing (14).

Finally, to determine the appropriate values of Q and λ, we use a cross-validation technique as described in Section 3.2.

3.2 Prediction ability of the model

For practical purposes, the final model is obtained by choosing simultaneously the optimal number of classes Q and the penalty parameter λ through a validation technique such as cross-validation (Stone, 1974). Specifically, similarly to Tutz and Ulbricht (2009), we refer to a cross-validation that consists of splitting the data set into three subsets: the training set, which is used to estimate the model parameters, the validation set to select λ and Q, and the test set to compute the prediction ability of the model. For each Q and λ ($Q = Q_1, .., Q_r$, $\lambda = \lambda_1, ..., \lambda_h$, where r and h are the different values of Q's and λ's considered, respectively), the model is fitted on the training set only, obtaining the estimates $\hat{\mathbf{W}}^Q_\lambda$, $\hat{\mathbf{V}}^Q_\lambda$, $\hat{\mathbf{C}}^Q_\lambda$, corresponding to the minimum value of (14) over 100 random starting points. Then, we compute the Mean Square Error $MSE^{(Q)}_\lambda$, on the validation set of size n_{val} as follows:

$$MSE^{(Q)}_\lambda = \| \mathbf{Y}_{val} - \mathbf{X}_{val}\hat{\mathbf{W}}^Q_\lambda\hat{\mathbf{V}}^Q_\lambda\hat{\mathbf{C}}^Q_\lambda \|^2 \frac{1}{n_{val}M}.$$

In this way, we build a grid of $MSE^{(Q)}_\lambda$ values tuning λ and Q simultaneously. Thus, we select the smallest λ^* and Q^* for which $MSE^{(Q)}_\lambda$ is minimum and we can assess the model performance by using the Mean Square Error ($MSE^{(Q^*)}_{\lambda^*}$) on the test set of size n_{test}, as follows:

$$MSE^{(Q^*)}_{\lambda^*} = \| \mathbf{Y}_{test} - \mathbf{X}_{test}\hat{\mathbf{W}}^*\hat{\mathbf{V}}^*\hat{\mathbf{C}}^* \| \frac{1}{n_{test}M},$$

where $\hat{\mathbf{W}}^*$, $\hat{\mathbf{V}}^*$ and $\hat{\mathbf{C}}^*$ are the estimates obtained on the training set with λ^* and Q^*.

Once predictors are partitioned into classes, a path diagram can be used to graphically display the relationships among the variables. Without loss of generality, as shown in Figure 1, predictors x_i are clustered into Q classes and connected to only one latent factor z_q $(q = 1, ..., Q)$, which in turn is connected to several responses $(y_1, ..., y_M)$. Relations between variables are indicated by lines and the lack of a line indicates no relationship. A line with one arrow represents a directed relationship between two variables, where the arrow points toward the dependent variable. Dotted lines indicate weak relationships between latent factors.

3.2.1 Illustrative example

We now show an example to better describe the penalized procedure, which is crucial in MRBOP estimation. Let us consider a simulated data set with $M = 2$ responses, J=10 predictors partitioned into 3 classes (the first three predictors in class 1, the second three in class 2 and the remaining four in class 3): the pairwise correlations between predictors in the same class have been set to 0.90 and the pairwise correlations between predictors in different classes have been set to 0.04 (Figure 2). Moreover, we have set

$$\mathbf{C} = \begin{pmatrix} 0.19 & 0.28 \\ 0.29 & 0.35 \\ 0.48 & 0.42 \end{pmatrix},$$

$\mathbf{w} = (0.49, 0.57, 0.66, -0.57, -0.77, -0.29, 0.40, 0.56, 0.65, 0.32)'$, I =800 and $\mathbf{E} \sim MVN(\mathbf{0}, \mathbf{I})$.
When model (6) is fitted with no penalty and $Q = 3$, the three variables in the true class 1 are allocated to three different groups, while the other predictors are clustered together in class 1. The correlations between the corresponding latent factors are respectively: $\text{corr}(\mathbf{z}_1, \mathbf{z}_2) = 0.14$, $\text{corr}(\mathbf{z}_1, \mathbf{z}_3) = 0.85$ and $\text{corr}(\mathbf{z}_2, \mathbf{z}_3) = 0.92$. It seems that the model tries to explain the largest part

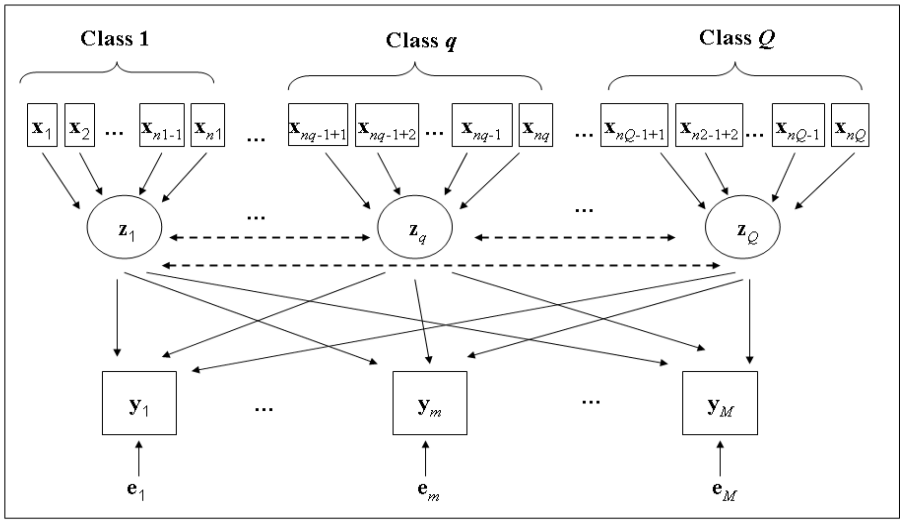

Fig. 1: Path diagram of MRBOP model

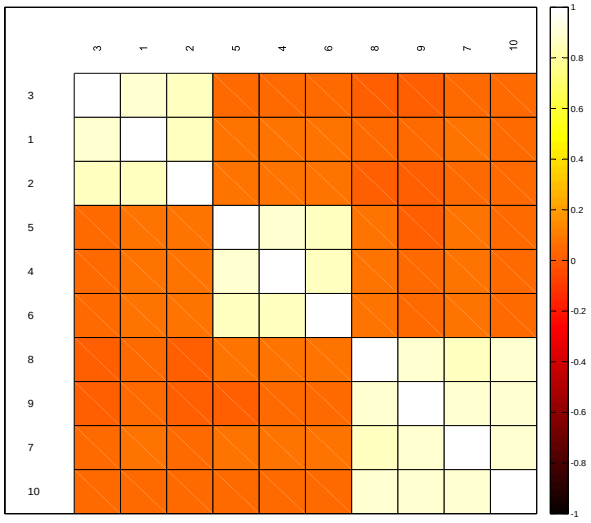

Fig. 2: Heat map of the correlation matrix of the simulated data set

of the total variance of the data by using a single latent factor $\mathbf{z}_1$, and creating two redundant correlated classes (classes 2 and 3) to predict the responses. Moreover, the model does not care about the relationships among predictors within the same class (they could be strongly or weakly correlated). Finally, the penalized MRBOP model has been fitted by using the cross-validation procedure as in Section 3.2 run on a grid tuning λ and Q simultaneously (the best solution retained over 100 random starting partitions). The training, validation, test sets were 85%, 10%, 5% of the whole sample, respectively, $Q \in \{2, 3, 4, 5, 6\}$ and $\lambda \in [0.1, 1.6]$ with increment of 0.10. Table 1 displays the MSE values from the test sets (for brevity, results for greater λ are not shown).

λ	$Q=2$	$Q=3$	$Q=4$	$Q=5$	$Q=6$
0.1	0.0567	0.0529	0.0525	0.0526	0.0525
0.2	0.0567	0.0528	0.0525	0.0525	0.0525
0.3	0.0528	0.0528	0.0525	0.0525	0.0525
0.4	0.0528	**0.0525**	0.0525	0.0525	0.0525
0.5	0.0528	0.0525	0.0525	0.0525	0.0525
0.6	0.0528	0.0525	0.0525	0.0525	0.0525
0.7	0.0528	0.0525	0.0525	0.0525	0.0525
0.8	0.0528	0.0525	0.0525	0.0525	0.0525
0.9	0.0528	0.0525	0.0525	0.0526	0.0525
1.0	0.0528	0.0525	0.0526	0.0525	0.0525
1.1	0.0528	0.0525	0.0525	0.0525	0.0525
1.2	0.0528	0.0525	0.0525	0.0525	0.0525
1.3	0.0528	0.0525	0.0525	0.0526	0.0525
1.4	0.0528	0.0525	0.0525	0.0526	0.0525
1.5	0.0528	0.0525	0.0525	0.0526	0.0525
1.6	0.0528	0.0525	0.0525	0.0526	0.0525

Table 1: MSE for (λ, Q) from the cross-validation procedure on the illustrative simulated data set (best choice in bold)

As evident, the best choice for λ and Q corresponds to the smallest values of λ and Q where MSE is minimum in the table, i.e. $\lambda^* = 0.4$ and $Q^* = 3$. Such a solution is able to recover the block structure of the predictors

with correlations between the three latent factors equal to -0.004, 0.001 and 0.005, respectively.

4 Simulation studies

In this section, we have evaluated the performance of the MRBOP model compared to OLS, PCR and PLSR on a set of 27 simulated data sets. The underlying true regression model is given by (6) with $M = 2$ and $\mathbf{E} \sim MVN(\mathbf{0}, d^2\mathbf{I})$, $(d = 0.1, 1, 2)$ where the constant d allows for different error levels. In the simulations all variables have been centered and standardized. Following the cross-validation procedure described in Section 3.2, each simulated data set is splitted into three subsets: $\mathbf{X}_{train}$, $\mathbf{X}_{val}$ and $\mathbf{X}_{test}$ of sizes $n_{train}|n_{val}|n_{test}$, respectively. Let $\bar{\mathbf{X}}_{train} = (\bar{x}_{1,train}, ..., \bar{x}_{J,train})'$ denote the $(J \times 1)$ mean vector of the predictors and $\bar{\mathbf{Y}}_{train} = (\bar{\mathbf{y}}_{1,train}, ..., \bar{\mathbf{y}}_{M,train})'$ the $(M \times 1)$ mean vector of the responses in the training set. Thus, the model is fitted on the training set only by setting $Q = 3$, retaining the best solution over 100 different starts and obtaining the estimates $\hat{\mathbf{W}}_\lambda$, $\hat{\mathbf{V}}_\lambda$, $\hat{\mathbf{C}}_\lambda$ for each $\lambda \in (0, 10]$ with increment of 0.10. Then, the corresponding MSE_λ are computed on the validation set as follows:

$$MSE_\lambda = \| (\mathbf{X}_{val}\mathbf{WVC} - \bar{\mathbf{Y}}_{train} + $$

$$+(\mathbf{X}_{val} - \bar{\mathbf{X}}_{train})\hat{\mathbf{W}}\hat{\mathbf{V}}\hat{\mathbf{C}}) \|^2 \frac{1}{n_{val}M}$$

and the smallest λ^* for which MSE_λ is minimum is selected. By using $\hat{\mathbf{W}}^*$, $\hat{\mathbf{V}}^*$ and $\hat{\mathbf{C}}^*$ corresponding to λ^*, the model performance is measured by:

- The test error,

$$MSE_{\lambda^*} = \| (\mathbf{X}_{test}\mathbf{WVC} - \bar{\mathbf{Y}}_{train} + $$
$$+ (\mathbf{X}_{test} - \bar{\mathbf{X}}_{train})\hat{\mathbf{W}}^*\hat{\mathbf{V}}^*\hat{\mathbf{C}}^*) \|^2 \frac{1}{n_{test}M};$$

- The L_2-distance between the true and the estimated coefficients to evaluate the accuracy of the estimators,

$$L_2 = \| \mathbf{WVC} - \hat{\mathbf{W}}^*\hat{\mathbf{V}}^*\hat{\mathbf{C}}^* \|^2;$$

- The Modified Rand Index (MRand) (Hubert and Arabie, 1985) which measures the degree of agreement between the true and the estimated partitions, MRand =

$$= \frac{\binom{J}{2}(a+f) - [(a+b)(a+c)+(c+f)(b+f)]}{\binom{J}{2}^2 - [(a+b)(a+c)+(c+f)(b+f)]}$$

  where
  - a - pairs of variables that are in the same class in $\mathbf{V}$ and in the same class in $\hat{\mathbf{V}}^*$;
  - b - pairs of variables that are in the same class in $\mathbf{V}$ and in different classes in $\hat{\mathbf{V}}^*$;
  - c - pairs of variables that are in the same class in $\hat{\mathbf{V}}^*$ and in different classes in $\mathbf{V}$;
  - f - pairs of variables that are in different classes in both $\hat{\mathbf{V}}^*$ and $\mathbf{V}$.

  The index is equal to 1 in case of perfect agreement.
- Percentage of successes to recover the exact partition, i.e. MRand = 1 (in Table 2 it is indicated as %=1).

The procedure is repeated $B = 100$ times and the resulting measures of performance are averaged over the replications. The experimental design has been set as follows. Each data set is of 300|300|200 units, $J = 10$ predictors partitioned in $Q = 3$ groups of sizes 3,3,4, respectively, $\mathbf{w} = (0.49, 0.57, 0.65, -0.57, -0.77, -0.29, 0.40, 0.56, 0.65, 0.32)$, $\mathbf{C} = \begin{pmatrix} 0.24 & 0.35 \\ 0.37 & 0.45 \\ 0.62 & 0.53 \end{pmatrix}$, three error levels (low $d = 0.1$, medium $d = 1$, high $d = 2$), and correlation matrix given by

$$r_{jj'} = \begin{cases} \rho_{wit} + 0.01U(0,1) \text{ if } \quad (j,j') \in q \quad (j \neq j') \\ \rho_{bet} + 0.01U(0,1) \text{ otherwise} \end{cases},$$

where $q = 1, 2, 3$ and $U(0, 1)$ is a uniform distribution in [0,1], ρ_{bet} defines the correlation between predictors in different groups, while ρ_{wit} is the correlation between predictors in the same group. Four different settings have been considered:

1. Setting 1: $\rho_{wit} = 0.90$, $\rho_{bet} = \{0.10, 0.30, 0.60\}$. This setting generates 9 experimental cells by crossing the three error levels with three correlation levels for ρ_{bet}.
2. Setting 2: $\rho_{wit} = 0.70$, $\rho_{bet} = \{0.10, 0.30, 0.60\}$, so that 9 experimental cells are generated.
3. Setting 3: $\rho_{wit} = 0.50$, $\rho_{bet} = \{0.10, 0.30\}$. This setting generates 6 experimental cells.
4. Setting 4: $\rho_{wit} = 0.30$, $\rho_{bet} = 0.10$. This setting generates 3 experimental cells.

Thus, in each setting $\mathbf{X}$ is drawn from a multivariate Normal distribution with mean vector $\mathbf{0}$ and correlation matrix $\mathbf{R} = [r_{jj'}]$. The simulation results are reported in Table 2 where the best performance is given in boldface. In all settings the increasing error level does not affect considerably the recovery of the true partition of predictors. In fact, when d increases, the average MRand values and the percentages of successes in recovering the true partition remain quite stable and exhibit acceptable values, even in the worst cases corresponding to settings 2 and 4, where the block correlation matrix generated is not well-defined (ρ_{wit} is

not much larger than ρ_{bet}). In particular, in setting 2 with ρ_{wit}=0.70 and ρ_{bet}=0.60, the average MRand is 0.79 and the algorithm is still able to recover the true structure of the predictors 52 times (out of 100). On the other hand, as expected, the error level influences the accuracy of the estimates in all settings: the higher the values of d, the higher the values of average L_2. Moreover, the error affects more the accuracy when the correlations within groups are strong, i.e. even though the average L_2 shows better results in setting 1 than in setting 4 for low error, when the error increases, the L_2 values get worse faster in setting 1 than in setting 4. The error level, instead, has a weak influence on the prediction performance: the increasing of d does not meaningfully affect the MSE values and its performance is slightly worse in setting 4, where the correlations within and between groups are both weak (ρ_{with} = 0.30 ρ_{bet} = 0.10). The algorithm converges quickly in a few iterations even in the worst case (i.e. 6 on average). Finally, it can be noticed that even when the error level is high, the general performance of the algorithm is good for all settings (i.e. low values of MSE and L_2) and MRBOP outperforms both PCR and PLSR. In particular, we can observe that OLS results are similar to the ones from MRBOP; this assures that the proposed model not only leads to a dimensionality reduction of the problem but also guarantees an optimal prediction and estimation accuracy. The three dimensionality reduction methods (MRBOP, PCR, PLSR) have been applied given the true number ($Q = 3$) of latent factors, to better compare the performance of the prediction. Obviously, a greater number of latent factors both for PCR and PLSR could have resulted in a better fit, but at the expense of the parsimony in terms of dimensionality reduction.

5 Application to Epidemiological data

The MRBOP model has been applied to epidemiological data in order to predict variables related to animal health from variables related to breeding environment, alimentary factors and farm management, among others. It is an important analysis which allows breeders to reduce disease at multiple points in the transmission cycle. The data set (Chauvin et al., 2005) consists of the measurements on $I = 659$ turkey flocks and $M = 2$ variables related to animal health ($\mathbf{Y}$): the farmer loss in terms of mortality ($\mathbf{y}_1$) and condemnation at slaughterhouse ($\mathbf{y}_2$). The $J = 19$ predictors (risk factors) are organized into 2 blocks: $\mathbf{X}_1$ and $\mathbf{X}_2$, where $\mathbf{X}_1$ includes 14 variables pertaining to farming features, while $\mathbf{X}_2$ includes 5 variables referred to technical and economical results of the flocks (variable description in Table 3). Since all variables are measured on different scales, they have been centered and scaled to unit variance. The data set has already been investigated by Bougeard et al. (2007) and Bougeard et al. (2008). In particular, the multiblock latent root regression proposed by Bougeard et al. (2007) shows that the most variance is accounted for by two out of five latent selected components. Such components are formed by all the predictors with different weights. In particular, the first component is mainly characterized by MOYVET, DENSI, VET, DWG, RI, KGM2 and TCI and the second one by ECI and TCI. Moreover, in their analysis the predictors that have the most influence on COMDEMN are SANPB, MOYVET and VET while MORT is mainly predicted by ECI, TCI and RI. On the other hand, Bougeard et al. (2008) compared the results of Redundancy Analysis (RA) and PLS regression. They also focus only on the first two latent components and show that only a few predictors (RI and DWG) are related to the first RA component while PLS regression manages to explain most of the predictors. RA achieves a better fitting of the responses, especially CONDEMN, whereas the latent components from PLS regression seem to be more linked to the predictors. Actually, CONDEMN is strongly related to the first component while MORT is best predicted by the second one. Critical to both approaches is the interpretability of the latent components which are not exclusively defined by only one set of predictors - for example TCI characterizes both components in Bougeard et al. (2007).

We have fitted a MRBOP model for different values of Q (from 1 to 10) and λ (0.1-10, with increments of 0.10), by using the cross-validation procedure described in Section 3.2, (training, validation and test sets were 85%, 10%, 5% of the whole sample, respectively) and 100 different starting points to avoid local minima. The algorithm converged in 9 iterations leading to the choice of three latent factors (corresponding to $MSE^{(Q^*=3)}_{\lambda^*=0.9} = 0.5933$), which are weakly correlated (0.174, 0.003, 0.244, respectively) as displayed in the path diagram of the estimated model (Figure 3). It turns out that the latent factor $\mathbf{z}_1$ is formed by the economical and technical results of the flocks (except for ECI) together with MOYVET, REMOV, OTHSPEC, DENSI, VET, strongly correlated to the economical aspects and it is mainly characterized by the variables DWG, RI and TCI. Latent factor $\mathbf{z}_2$ mainly reflects the remaining farming characteristics as biosecurity (through the variable DISINF), surface availability (SURF) and other general features related to the farms as FEED, PAREMPTY, LITTER and ECI (the largest normalized weights are in boldface in Figure 3). Finally, $\mathbf{z}_3$ corresponds to one single binary

Setting 1		$d=0.1$				$d=1$				$d=2$			
$\rho_{wit}=0.90$	Model	MSE_{λ^*}	L_2	MRand	%=1	MSE_{λ^*}	L_2	MRand	%=1	MSE_{λ^*}	L_2	MRand	%=1
$\rho_{bet}=0.10$	MRBOP	**0.0012**	**0.0010**	1.00	99	**0.0012**	**0.0019**	1.00	99	**0.0014**	**0.0046**	0.98	95
	Ols	**0.0012**	**0.0010**	-	-	0.0013	0.0028	-	-	0.0015	0.0079	-	-
	PCR	0.0033	0.0495	-	-	0.0033	0.0499	-	-	0.0034	0.0488	-	-
	PLSR	0.0031	0.0442	-	-	0.0031	0.0447	-	-	0.0031	0.0436	-	-
$\rho_{bet}=0.30$	MRBOP	**0.0012**	**0.0011**	1.00	99	**0.0011**	**0.0020**	1.00	99	**0.0012**	**0.0052**	0.97	93
	Ols	**0.0012**	**0.0011**	-	-	0.0012	0.0030	-	-	0.0014	0.0089	-	-
	PCR	0.0035	0.0566	-	-	0.0036	0.0559	-	-	0.0035	0.0562	-	-
	PLSR	0.0032	0.0498	-	-	0.0033	0.0490	-	-	0.0031	0.0491	-	-
$\rho_{bet}=0.60$	MRBOP	**0.0006**	0.0016	0.96	88	**0.0008**	**0.0031**	0.97	94	**0.0012**	**0.0079**	0.92	82
	Ols	**0.0006**	**0.0010**	-	-	**0.0008**	0.0035	-	-	0.0013	0.0108	-	-
	PCR	0.0037	0.0719	-	-	0.0039	0.0727	-	-	0.0040	0.0708	-	-
	PLSR	0.0030	0.0567	-	-	0.0033	0.0579	-	-	0.0033	0.0562	-	-
Setting 2		$d=0.1$				$d=1$				$d=2$			
$\rho_{wit}=0.70$	Model	MSE_{λ^*}	L_2	MRand	%=1	MSE_{λ^*}	L_2	MRand	%=1	MSE_{λ^*}	L_2	MRand	%=1
$\rho_{bet}=0.10$	MRBOP	0.0011	0.0015	1.00	98	**0.0012**	**0.0018**	1.00	98	**0.0012**	**0.0026**	0.96	91
	Ols	**0.0010**	**0.0011**	-	-	**0.0012**	0.0019	-	-	0.0013	0.0038	-	-
	PCR	0.0093	0.0599	-	-	0.0098	0.0607	-	-	0.0094	0.0592	-	-
	PLSR	0.0063	0.0387	-	-	0.0067	0.0392	-	-	0.0064	0.0386	-	-
$\rho_{bet}=0.30$	MRBOP	0.0010	0.0015	0.98	94	**0.0010**	**0.0019**	0.97	92	**0.0012**	**0.0031**	0.96	89
	Ols	**0.0009**	**0.0012**	-	-	**0.0010**	0.0020	-	-	0.0014	0.0043	-	-
	PCR	0.0114	0.0715	-	-	0.0113	0.0706	-	-	0.0117	0.0736	-	-
	PLSR	0.0067	0.0398	-	-	0.0066	0.0392	-	-	0.0068	0.0409	-	-
$\rho_{bet}=0.60$	MRBOP	0.0008	0.0033	0.79	52	0.0009	0.0039	0.82	57	0.0012	**0.0051**	0.80	55
	Ols	**0.0004**	**0.0013**	-	-	**0.0006**	**0.0023**	-	-	**0.0011**	**0.0051**	-	-
	PCR	0.0191	0.1255	-	-	0.0191	0.1238	-	-	0.0182	0.1177	-	-
	PLSR	0.0042	0.0258	-	-	0.0042	0.0255	-	-	0.0044	0.0264	-	-
Setting 3		$d=0.1$				$d=1$				$d=2$			
$\rho_{wit}=0.50$	Model	MSE_{λ^*}	L_2	MRand	%=1	MSE_{λ^*}	L_2	MRand	%=1	MSE_{λ^*}	L_2	MRand	%=1
$\rho_{bet}=0.10$	MRBOP	**0.0014**	0.0021	0.97	91	**0.0012**	0.0021	0.96	89	**0.0014**	**0.0026**	0.97	92
	Ols	**0.0014**	**0.0018**	-	-	**0.0012**	**0.0020**	-	-	0.0016	0.0033	-	-
	PCR	0.209	0.0800	-	-	0.0214	0.0821	-	-	0.0208	0.0798	-	-
	PLSR	0.0081	0.0284	-	-	0.0081	0.0293	-	-	0.0080	0.0289	-	-
$\rho_{bet}=0.30$	MRBOP	0.0012	0.0024	0.93	84	0.0013	0.0026	0.95	87	**0.0016**	**0.0039**	0.92	78
	Ols	**0.0009**	**0.0017**	-	-	**0.0012**	**0.0025**	-	-	**0.0016**	0.0040	-	-
	PCR	0.0286	0.1118	-	-	0.0283	0.1104	-	-	0.0273	0.1050	-	-
	PLSR	0.0070	0.0258	-	-	0.0070	0.0256	-	-	0.0066	0.0231	-	-
Setting 4		$d=0.1$				$d=1$				$d=2$			
$\rho_{wit}=0.30$	Model	MSE_{λ^*}	L_2	MRand	%=1	MSE_{λ^*}	L_2	MRand	%=1	MSE_{λ^*}	L_2	MRand	%=1
$\rho_{bet}=0.10$	MRBOP	0.0016	**0.0029**	0.92	79	0.0016	0.0029	0.91	79	**0.0020**	**0.0037**	0.94	83
	Ols	**0.0014**	**0.0023**	-	-	**0.0013**	**0.0024**	-	-	0.0021	0.0040	-	-
	PCR	0.0515	0.1420	-	-	0.0508	0.1402	-	-	0.0484	0.1335	-	-
	PLSR	0.0064	0.0165	-	-	0.0062	0.0162	-	-	0.0066	0.0168	-	-

Table 2: Simulation results

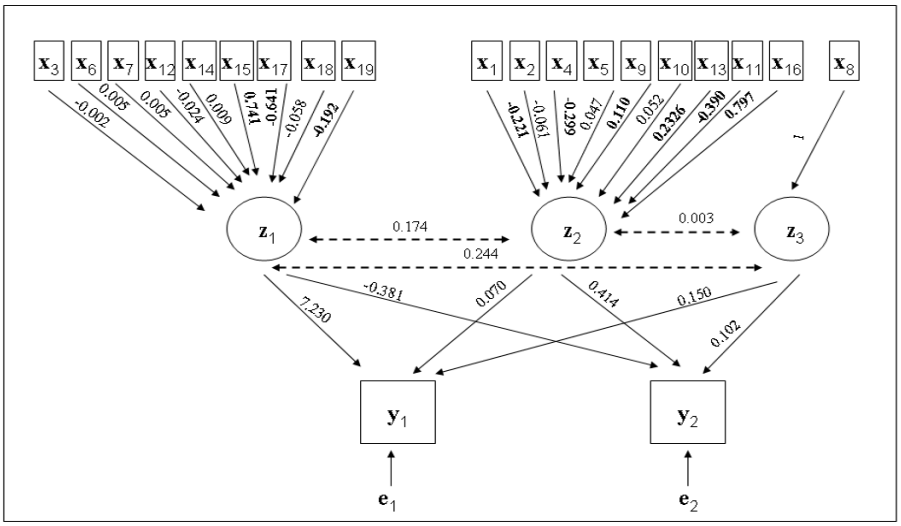

Fig. 3: Epidemiological data: path diagram

variable (SANPB), representing health problems during farming. Both responses (CONDEMN and MORT) are mainly explained by $\mathbf{z}_1$, even though for the carcasses condemnation (CONDEMN) the effect is much stronger (the corresponding coefficients are $c_{11}=7.230$ vs $c_{12}=-0.381$). CONDEMN is also influenced by $\mathbf{z}_3$ while MORT by $\mathbf{z}_2$. In other words, particular care for the technical performance of the flocks and the health problems during farming should be taken in order to reduce the number of carcasses condemned at slaughterhouse. On the other hand, to reduce the mortality (MORT), technical performance and some farm characteristics of the flocks need to be improved. Table 4 displays the regression coefficients of the single predictors (given by $\hat{\mathbf{W}}^*\hat{\mathbf{V}}^*\hat{\mathbf{C}}^*$) together with the 90% confidence intervals computed by a bootstrap procedure

Variables	ID	Description
$\mathbf{Y}$		
$\mathbf{y}_1$	CONDEMN	Percentage of carcasses condemned at slaughterhouse
$\mathbf{y}_2$	MORT	Mortality percentage for the flock
$\mathbf{X}_1$		
$\mathbf{x}_1$	FEED	Meat and bone meal-free feeding (1 = yes, 0 = no)
$\mathbf{x}_2$	COSTDIS	Disinfection costs
$\mathbf{x}_3$	MOYVET	Average veterinary costs for the last three flocks
$\mathbf{x}_4$	PAREMPTY	Partial emptying (1 = yes, 0 = no)
$\mathbf{x}_5$	DEMPTY	Duration of the empty period before chick arrival
$\mathbf{x}_6$	REMOV	Number of removal to slaughterhouse per flock
$\mathbf{x}_7$	OTHSPEC	Last flock with the same species (1 = yes, 0 = no)
$\mathbf{x}_8$	SANPB	Serious health problem during farming (1 = yes, 0 = no)
$\mathbf{x}_9$	DISINF	Disinfection labour (1 = skilled labour, 0 = yourself)
$\mathbf{x}_{10}$	CLEAN	Cleaning labour (1 = skilled labour, 0 = yourself)
$\mathbf{x}_{11}$	LITTER	Quantity of litter used for the flock
$\mathbf{x}_{12}$	DENSI	Chick density at the beginning of farming
$\mathbf{x}_{13}$	SURF	Surface area on which the flock is farmed
$\mathbf{x}_{14}$	VET	Total amount of veterinary costs for the flock
$\mathbf{X}_2$		
$\mathbf{x}_{15}$	DWG	Daily weight gain
$\mathbf{x}_{16}$	ECI	Economical consumption index
$\mathbf{x}_{17}$	RI	Result index
$\mathbf{x}_{18}$	KGM2	Total flock weight slaughtered related to the surface area
$\mathbf{x}_{19}$	TCI	Technical consumption index

Table 3: Epidemiological data: variable description (Chauvin et al., 2005)

VARIABLE	COMDEMN	LOWER LIMIT	UPPER LIMIT	MORT	LOWER LIMIT	UPPER LIMIT
FEED	-0.0155	-0.0789	0.0054	**-0.0916**	-0.1881	-0.0025
COSTDIS	-0.0042	-0.0588	0.0219	-0.0251	-0.0410	0.1021
MOYVET	-0.0144	-0.0663	0.1037	0.0008	-0.0040	0.2646
PAREMPTY	-0.0209	-0.0321	0.0321	-0.1238	-0.1436	0.0008
DEMPTY	0.0033	-0.0113	0.0641	0.0192	-0.0527	0.0465
REMOV	0.0351	-0.0420	0.0494	-0.0019	-0.2135	0.0018
OTHSPEC	0.0392	-0.0265	0.0818	-0.0021	-0.1176	0.0064
SANPB	**0.1504**	0.0764	0.1871	**0.1023**	0.0104	0.1771
DISINF	0.0077	-0.0345	0.0340	0.0456	-0.0125	0.1277
CLEAN	-0.0036	-0.0713	0.0043	-0.0215	-0.1184	0.0267
LITTER	**-0.0272**	-0.1229	-0.0024	-0.1610	-0.1899	0.0034
DENSI	-0.1722	-0.3481	0.0910	0.0091	-0.2345	0.0800
SURF	0.0163	-0.0135	0.0637	**0.0963**	0.0014	0.2250
VET	0.0678	-0.0829	0.1231	-0.0036	-0.3384	0.0070
DWG	**5.3549**	4.8340	5.7154	-0.2823	-0.4533	0.8100
ECI	0.0558	-0.0257	0.1374	**0.3299**	0.2626	0.4308
RI	**-4.6328**	-4.9680	-4.1939	0.2443	-0.6963	0.3990
KGM2	**-0.4211**	-0.6143	-0.0473	0.0222	-0.4329	0.0316
TCI	**-1.3865**	-1.5568	-1.0549	0.0731	-0.2036	0.1152

Table 4: Epidemiological data: MRBOP regression coefficients and 90% bootstrap confidence intervals

with 1000 bootstrap samples. Looking more in details at the regression coefficient estimates and their corresponding confidence intervals, we can observe that the percentage of carcasses condemned at slaughterhouse is influenced mainly by the technical results of the flocks (DWG, RI, TCI, KGM2), presence of health problem (SANPB) and quantity of litter (LITTER). As far as MORT concerns, FEED has a significant negative effect, while ECI, SANPB and SURF have significant positive coefficients. We can conclude that a reduction of the percentage of carcasses can be achieved mainly by improving technical results of the flocks, while in order to reduce the mortality percentage, some farming (mainly related to sanitary and environmental) features and economical results of the flocks need to be controlled.

6 Conclusions

In this paper a new model MRBOP for multivariate regression based on a small set of weakly correlated latent factors is presented, where each of the latent factors is a linear combination of a subset of correlated predictors. MRBOP is particularly appropriate in a regression context where strongly correlated predictors might represent unknown underlying latent dimensions easy to be interpreted. The performance of the proposed approach has been discussed on both simulated and real data sets and generally exhibits accuracy of the estimates and capability to recover the block correlation structure of the original predictors. As suggested by one referee, further developments could extend the linear factor regression based on an "ignoring errors" strategy (since the loss function does not include the error matrix) to an approach able to fit the model with the errors being part of the loss function.

Acknowledgements The authors are grateful to the editor and anonymous referees of Statistics and Computing for their valuable comments and suggestions which improved the clarity and the relevance of the first version.

References

Abdi, H.: Partial least squares regression and projection on latent structure regression (PLS-Regression), Wiley Interdisciplinary Reviews: Computational Statistics, **2**, 97–106 (2010)

Abraham, B., Merola, G.: Dimensionality reduction approach to multivariate prediction, Computational Statistics & Data Analysis, **48**(1), 516 (2005)

Anderson, T.W.: Estimating linear restrictions on regression coefficients for multivariate distributions, Annals of Mathematical Statistics, **22**, 327–351 (1951)

Bougeard, S., Hanafi, M., Qannari, E.M.: Multiblock latent root regression. Application to epidemiological data, Computational Statistics, **22**(2), 209–222 (2007)

Bougeard, S., Hanafi, M., Qannari, E.M.: Continuum redundancy-PLS regression: A simple continuum approach, Computational Statistics and Data Analysis, **52**(7), 3686–3696 (2008)

Chauvin, C., Buuvrel, I., Belceil, P.A., Orand, J.P., Guillemot, D., Sanders, P.: A pharmaco-epidemiological analysis of factors associated with antimicrobial consumption level in turkey broiler flocks, **36**, 199–211 (2005)

De Jong, S.: SIMPLS: an alternative approach to partial least squares regression, Chemometrics and Intelligent Laboratory Systems, **18**, 251-263 (1993)

Escoufier, Y.: Le traitement des variables vectorielles, Biometrics, **29**, 751–760 (1973)

Frank, I.E., Friedman, J.: A statistical view of some chemometrics regression tools, Technometrics, **35**, 109–148 (1993)

Hocking, R.R.: The Analysis and Selection of Variables in Linear Regression, Biometrics, **32**, 1–49 (1976)

Hoerl, A.E., Kennard, R.W.: Ridge Regression: Biased Estimation for Non-Orthogonal Problems, Technometrics, **12**, 55–67 (1970)

Hotelling, H.: The most predictable criterion, Journal of Educational Psychology, **25**, 139–142 (1935)

Hubert, L., Arabie, P.: Comparing partitions, Journal of Classification, **198**, 193-218 (1985)

Izenman, A.J.: Reduced-Rank Regression for the Multivariate Linear Model, Journal of Multivariate Analysis, **5**, 248–262 (1975)

Jolliffe, I.T.: A note on the Use of Principal Components in Regression, Journal of the Royal Statistical Society: Series C (Applied Statistics), **31**(3), 300–303 (1982)

Krzanowski, W.,J.: Principles of Multivariate Analysis: A User's Perspective, Oxford University Press, (2000)

Rosipal, R., Krmer, N.: Overview and Recent Advances in Partial Least Squares, In Subspace, Latent Structure and Feature Selection, **3940**, 34–51 (2006)

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

Stone, M., Brooks, R.J.: Continuum Regression: Cross-validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression, Journal of the Royal Statistical Society: Series B, **52**(2), 237–269, (1990)

Tibshirani, R.: Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society Series B, **58**(1), 267–288 (1996)

Tutz, G., Ulbricht, J.: Penalized regression with correlation-based penalty, Statistics and Computing, **19**(1), 239–253 (2009)

Van Den Wollenberg, A.L.: Redundancy analysis an alternative for canonical correlation analysis, Psychometrika, **42**(2), 207–219 (1977)

Waldro, L., Pintilie, M., Tsao, M.S., Shepherd, F.A., Huttenhower, C., Jurisica, I.: Optimized application of penalized regression methods to diverse genomic data, Bioinformatics, **27**(24), 3399–3406 (2011)

Witten, D.M., Tibshirani, R.: Covariance-regularized regression and classification for high dimensional problems, Journal of the Royal Statistical Society: Series B (Statistical Methodology), **71**, 615–636 (2009)

Wold, H.: Estimation of principal components and related models by iterative least squares, In: P.R. Krishnaiaah (Ed.), Multivariate Analysis, New York, 391–420 (1966)

Yuan, M., Ekici, A., Lu, Z., Monteiro, R.: Dimension reduction and coefficient estimation in multivariate linear regression, Journal of the Royal Statistical Society: Series B (Statistical Methodology), **69**, 329–346 (2007)

Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society: Series B, **68**, 49–67 (2006)

Zou, H., Hastie, T.: Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society Series B, **67**, 301–320 (2005)