Chemometrics and Intelligent Laboratory Systems 123 (2013) 36–43

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

On the selection of the weighting parameter value in Principal Covariates Regression

Marlies Vervloet *, Katrijn Van Deun, Wim Van den Noortgate, Eva Ceulemans

KU Leuven, Andreas Vesaliusstraat 2, Bus 3762, 3000 Leuven, Belgium

ARTICLE INFO

Article history:
Received 16 October 2012
Received in revised form 8 February 2013
Accepted 15 February 2013
Available online 24 February 2013

Keywords:
Principal Covariates Regression
Regression models
Dimension reduction
Weighting

ABSTRACT

Ordinary linear regression falls short when many predictors are available, especially when some of these are highly correlated with (a linear combination of) other predictors. One possible solution for this problem is Principal Covariates Regression (PCovR) which combines the main ideas behind Principal Component Analysis (PCA) and regression. Like PCA, PCovR reduces the predictors to a few components and, like regression, it predicts the criterion, but using the components as predictors. The reduction of the predictors and the prediction of the criterion are conducted simultaneously, by minimizing the weighted sum of the reduction error and the prediction error. How the value of the weighting parameter α can be optimally tuned, is not so obvious however. In this paper we integrate scattered findings on this topic and derive some hypotheses on which α value is optimal in which respect and on the importance of tuning the value (how robust are the obtained results for the value that is chosen for α?). We put these hypotheses to the test by performing an extensive simulation study. As predicted, the α value that optimizes recovery of the underlying parameters and true criterion scores depends among others on the number of predictors in a specific dataset and on the ratio of the amount of error on the predictors and the amount of error on the criterion. Moreover, we show that α is mostly of influence when the components strongly differ in strength and relevance, when the number of observations almost equals the number of predictor variables, or when the criterion contains a moderate to high amount of error.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

In analytical chemistry, ordinary least squares regression (OLS) is often used to predict the scores of N observations on a criterion by means of their scores on J predictors. Bro, Nielsen, Stefánsson and Skåra [1], for example, examine whether the ripening of several samples of salted herring as assessed by a trained panel, can be predicted on the basis of chemical attributes of the samples. OLS may become troublesome, however, when there are a lot of predictor variables, because each regression weight reveals the additional contribution of the predictor on top of all other predictors and therefore all regression weights should be interpreted simultaneously. To complicate things further, the chances get higher that at least some of the predictor variables will be highly correlated with a linear combination of one or more other predictor variables. Such (multi)collinearity creates a supplementary problem, called the bouncing beta problem, which implies that small differences in the data, such as removing or adding one single observation, can create huge differences in the regression weights [2].

Many solutions to the bouncing beta problem exist, among which the penalty-based approaches (e.g., Lasso [3]) are very popular. A drawback of these approaches, in which some of the regression weights are shrunken to zero, is that they typically do not yield insight into the correlation structure of the predictors. Therefore, in this paper, we focus on another type of solution, that is, linear dimension reduction. This approach implies that the predictor variables are reduced to R $(1 \leq R < J)$ new variables, called components, which are linear combinations of the predictors:

$$\mathbf{T} = \mathbf{XW}, \tag{1}$$

where $\mathbf{T}$ is an $N \times R$ matrix that contains the scores of every observation on the R components, $\mathbf{X}$ is the $N \times J$ predictor data matrix, and $\mathbf{W}$ is a $J \times R$ weight matrix. The criterion $\mathbf{y}$ $(N \times 1)$ is regressed on the components instead of on the predictors:

$$\mathbf{y} = \mathbf{Tp_y} + \mathbf{e_y}, \tag{2}$$

with $\mathbf{p_y}$ $(R \times 1)$ containing the resulting regression weights and $\mathbf{e_y}$ being the residual $\mathbf{y}$ scores. Such dimension reduction methods solve the interpretational problems of OLS, because fewer regression

* Corresponding author. Tel.: +32 16 3 26192.
E-mail addresses: marlies.vervloet@ppw.kuleuven.be (M. Vervloet), katrijn.vandeun@ppw.kuleuven.be (K. Van Deun), wim.vandennoortgate@kuleuven-kulak.be (W. Van den Noortgate), eva.ceulemans@ppw.kuleuven.be (E. Ceulemans).

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.02.005

weights have to be interpreted (if $R<J$) and because, through $\mathbf{W}$, insight is obtained into the correlation structure of the predictors. A number of such dimension reduction methods has been proposed which differ among others in the extent to which the summary of the predictors (1) or the prediction of the criterion (2) is emphasized. For instance, in reduced rank regression (RRR [4]) the obtained components predict the criterion as well as possible, but they are not necessarily good summarizers of the predictors, whereas in principal components regression (PCR [5]) one sequentially reduces the predictors to components through standard Principal Component Analysis (PCA) and uses those components for the prediction of the criterion, an approach that guarantees an optimal summary of the predictors, but not a good prediction of the criterion.

A method which allows to flexibly emphasize both summarizing and predicting, is Principal Covariates Regression (PCovR [6]). Using PCovR, the reduction of the predictors and the prediction of the criterion are done simultaneously, by minimizing the weighted sum of the reduction error and the prediction error:

$$min\left[\alpha\frac{||\mathbf{X}-\mathbf{T}\mathbf{P}_\mathbf{X}||^2}{||\mathbf{X}||^2}+(1-\alpha)\frac{||\mathbf{y}-\mathbf{T}\mathbf{p}_\mathbf{y}||^2}{||\mathbf{y}||^2}\right], \tag{3}$$

with $||\mathbf{X}||$ being the Frobenius matrix norm of $\mathbf{X}$. As in standard PCA, the loading matrix $\mathbf{P}_\mathbf{X}$ ($R\times J$) contains the loadings of each predictor variable on each component. These loadings express the correlation between the predictors and the components in case of orthogonal and standardized components and standardized predictors. α is the weighting parameter, ranging between 0 and 1, which determines to what degree the focus is put on the reduction of $\mathbf{X}$ rather than on the prediction of $\mathbf{y}$. Specifically, α values of 0 and 1 correspond, respectively, with RRR [4] and PCR [5]. Given a specific α value and number of components R, a closed form solution exists for estimating the PCovR parameters, by setting $\mathbf{T}$ equal to the first R normalized eigenvectors of the matrix

$$\mathbf{G}=\alpha\frac{\mathbf{X}\mathbf{X}'}{||\mathbf{X}||^2}+(1-\alpha)\frac{\mathbf{H}_\mathbf{X}\mathbf{y}\mathbf{y}'\mathbf{H}_\mathbf{X}}{||\mathbf{y}||^2}, \tag{4}$$

in which $\mathbf{H}_\mathbf{x}$ is the projection matrix which projects any vector on $\mathbf{X}$ [6].

Note that the PCovR loss function (3) can be rewritten as:

$$\begin{aligned} &min\left[\frac{\alpha}{||\mathbf{X}||^2}||\mathbf{X}-\mathbf{T}\mathbf{P}_\mathbf{X}||^2+\frac{(1-\alpha)}{||\mathbf{y}||^2}||\mathbf{y}-\mathbf{T}\mathbf{p}_\mathbf{y}||^2\right] \\ &=min\left[\beta_1||\mathbf{X}-\mathbf{T}\mathbf{P}_\mathbf{X}||^2+\beta_2||\mathbf{y}-\mathbf{T}\mathbf{p}_\mathbf{y}||^2\right]. \end{aligned} \tag{5}$$

Rescaling (5), by dividing β_1 and β_2 by $\beta_1+\beta_2$, does not change the obtained solution, but alters the value of the loss function, this function can be formulated as

$$min\left[\beta||\mathbf{X}-\mathbf{T}\mathbf{P}_\mathbf{X}||^2+(1-\beta)||\mathbf{y}-\mathbf{T}\mathbf{p}_\mathbf{y}||^2\right], \tag{6}$$

where β can be computed as:

$$\beta=\frac{\alpha||\mathbf{y}||^2}{\alpha||\mathbf{y}||^2+(1-\alpha)||\mathbf{X}||^2}. \tag{7}$$

Taking the interpretational advantages described above into account, as well as the fact that a closed form solution exists for estimating the parameters, we conclude that PCovR is potentially very useful. Yet, PCovR is not commonly used. A first reason is that the procedure is not well known and not implemented in standard statistical software. Another reason might be that researchers find the weighting parameter α difficult to handle in that they do not know how they have to select a particular value. A number of researchers have devoted some attention to this problem, but from different angles and in different application domains. Therefore, in this paper we will integrate these scattered findings and derive some hypotheses on the optimal α value and on the importance of tuning the value (how robust are the obtained results for the value that is chosen for α?) (Section 2). We will also put these hypotheses to the test by performing an extensive simulation study (Section 3). Finally, we formulate some conclusions.

2. Literature overview

Two strategies have already been used to determine the α value that optimizes the performance of PCovR, on the basis of characteristics of the data. The first strategy makes use of a stochastic extension of PCovR, by making assumptions about the error distribution and analyzing the loglikelihood of the data, while the other strategy is to simulate data and manipulate the α value to see which value leads to the most accurate estimates. At this point, we already want to emphasize that both strategies thus define optimality in a different way.

2.1. Stochastic approach

The stochastic approach, suggested by Gurden [8] and Wilderjans, Ceulemans and Van Mechelen [9], searches for an α value that maximizes the likelihood of the observed data, given particular PCovR estimates. This approach builds on the assumption that the error on the predictor block ($\mathbf{E}_\mathbf{X}$) and the error on the criterion block ($\mathbf{e}_\mathbf{y}$) follow normal distributions with mean 0 and variances $\sigma^2_{\mathbf{E}_\mathbf{X}}$ and $\sigma^2_{\mathbf{e}_\mathbf{y}}$. Given this assumption, maximum likelihood estimates can be obtained by minimizing the following loglikelihood $L\left(\mathbf{X},\mathbf{y}\middle|\mathbf{T},\mathbf{P}_\mathbf{x},\mathbf{p}_\mathbf{y},\sigma^2_{\mathbf{E}_\mathbf{X}},\sigma^2_{\mathbf{e}_\mathbf{y}}\right)$ of the data given the PCovR model as well as the two error variances

$$\begin{aligned} L\left(\mathbf{X},\mathbf{y}|\mathbf{T},\mathbf{P}_\mathbf{x},\mathbf{p}_\mathbf{y},\sigma^2_{\mathbf{E}_\mathbf{X}},\sigma^2_{\mathbf{e}_\mathbf{y}}\right)&=nJ\ln\left(2\pi\sigma_{\mathbf{E}_\mathbf{X}}\right)+n\ln\left(2\pi\sigma_{\mathbf{e}_\mathbf{y}}\right) \\ &\quad+\frac{||\mathbf{X}-\mathbf{T}\mathbf{P}_\mathbf{X}||^2}{2\sigma^2_{\mathbf{E}_\mathbf{X}}}+\frac{||\mathbf{y}-\mathbf{T}\mathbf{p}_\mathbf{y}||^2}{2\sigma^2_{\mathbf{e}_\mathbf{y}}}. \end{aligned} \tag{8}$$

Omitting the first two terms, as they are constant for a specific dataset, and noting that differences in block size and scaling are not taken into account, it can be concluded that this expression boils down to the β_1 and β_2 version of the PCovR loss function (5), with $\beta_1=\frac{1}{2\sigma^2_{\mathbf{E}_\mathbf{X}}}$ and $\beta_2=\frac{1}{2\sigma^2_{\mathbf{e}_\mathbf{y}}}$. Dividing β_1 and β_2 by $\beta_1+\beta_2$, we obtain the β version of the loss function (6), implying that maximum likelihood PCovR estimates are obtained for the following β value, indicated as β_{ML}:

$$\beta_{ML}=\frac{\beta_1}{\beta_1+\beta_2}=\frac{\frac{1}{2\sigma^2_{\mathbf{E}_\mathbf{X}}}}{\frac{1}{2\sigma^2_{\mathbf{E}_\mathbf{X}}}+\frac{1}{2\sigma^2_{\mathbf{e}_\mathbf{y}}}}=\frac{1}{1+\frac{\sigma^2_{\mathbf{E}_\mathbf{X}}}{\sigma^2_{\mathbf{e}_\mathbf{y}}}}. \tag{9}$$

Hence, maximum likelihood estimates of $\mathbf{T}$, $\mathbf{P}_\mathbf{X}$ and $\mathbf{p}_\mathbf{y}$ can be obtained by downweighting the more noisy data block in favor of the less noisy data block [10]. Given the relation between α and β (Eq. (7)) and assuming that each variable has equal sum of squares S, it can be derived that the corresponding α_{ML} value equals

$$\alpha_{ML}=\frac{||\mathbf{X}||^2}{||\mathbf{X}||^2+\frac{\sigma^2_{\mathbf{E}_\mathbf{X}}||\mathbf{y}||^2}{\sigma^2_{\mathbf{e}_\mathbf{y}}}}=\frac{JS}{JS+\frac{\sigma^2_{\mathbf{E}_\mathbf{X}}S}{\sigma^2_{\mathbf{e}_\mathbf{y}}}}=\frac{J}{J+\frac{\sigma^2_{\mathbf{E}_\mathbf{X}}}{\sigma^2_{\mathbf{e}_\mathbf{y}}}}. \tag{10}$$

It can be concluded that α_{ML} increases not only with a higher number of predictor variables, but also when the amount of error on $\mathbf{X}$ decreases relative to the amount of error on $\mathbf{y}$. Furthermore, the effect of this error ratio is expected to be more outspoken when fewer predictor variables are involved.

Table 1
Characteristics and results of four simulation studies.

	Gurden [8]	Wilderjans, Ceulemans & Van Mechelen [9]	Heij, Groenen & Van Dijk [11]	Kiers & Smilde [2]
Model	Multiway Covariates Regression	Multiway multiblock component model	PCovR	PCovR
Measure	Fit	Recovery	Fit	Recovery
Weighting parameter values used	$\alpha = 0, .01, .02 \ldots 1$	$\alpha = .50$ $\beta = .50$	$\alpha = .10, .50, .90, .99, .9999$	$\alpha = 0, .50, 1$
Optimal α	$.91 \leq \alpha \leq .98$	$\alpha > .50$	$.90 \leq \alpha \leq 1$	$\alpha > .50$
Optimal α influenced by	Amount of error	Amount of error, number of predictors	Amount of error, number of predictors	Amount of error, correlation between predictor variables

2.2. Simulation studies

With respect to the simulation approach, Heij, Groenen and Van Dijk [11] and Kiers and Smilde [2] examined the performance of PCovR by means of artificial data. Moreover, Wilderjans, Ceulemans and Van Mechelen [9] and Gurden [8] conducted a simulation study in which the influence of different weighting schemes on the performance of the multiway multiblock component model and the Multiway Covariates Regression model is investigated, which are both multiway extensions of PCovR. Note, however, that unlike Multiway Covariates Regression and PCovR, the multiway multiblock component model does not impose the restriction that the components are linear combinations of the predictors. As will be discussed in the next paragraphs, these studies yield insight into different aspects of PCovR performance, given that different data generation processes are used, different choices for the weighting parameter value are evaluated, and different performance measures are considered (see Table 1).

Regarding data generation, Heij, Groenen, and Van Dijk [11] created data according to the PCovR model. However, in most cases, only one component had a non-zero regression weight, implying that the other components are not relevant for predicting the criterion. Out of the five simulation studies that were performed by Kiers and Smilde [2], only one implied data generated on the basis of a PCovR model (with two components). However, the majority of the predictor variables had a very low correlation with the criterion variables as well as the components. The data of Wilderjans, Ceulemans and Van Mechelen [9] and Gurden [8] were multiway.

Regarding α values, Kiers and Smilde [2] only used values of 0, .50 and 1, when analyzing the data, and Wilderjans, Ceulemans and Van Mechelen [9] compared $\alpha = .50$ (which they call block-based weighting) with $\beta = .50$ (coined entry-based weighting). The most extensive manipulations were done by Heij, Groenen and Van Dijk [11] who considered $\alpha = .10, .50, .90, .99$ and .9999, and Gurden [8] who considered 101 α values, evenly spread between 0 and 1.

Finally, in the different studies different measures were used to evaluate PCovR performance, which mostly focus on how well the estimated model predicts other data, or on how well the underlying parameters are recovered. More specifically, Gurden [8] and Heij, Groenen and Van Dijk [11] investigate how well the model can predict future data, by performing cross-validation and calculating the mean squared forecast error, respectively. On the other hand, Wilderjans, Ceulemans and Van Mechelen [9] and Kiers and Smilde [2] inspect the goodness-of-recovery of all parameter sets (loadings and component scores) and of the regression weights of the predictor variables,[1] respectively.

If we turn to the conclusions of these simulation studies regarding the optimal α value (see Table 1), Kiers and Smilde [2] found that $\alpha = 1$, which boils down to applying PCR, always yields a better performance (i.e. better recovery of the regression weights and the true criterion scores) than $\alpha = 0$ (i.e., RRR) or $\alpha = .50$. This is in line with Wilderjans, Ceulemans and Van Mechelen [9], who report that $\beta = .50$ yields better results (i.e. better recovery of loadings and component scores) than $\alpha = .50$, as $\beta = .50$ indeed corresponds to a higher α value as soon as the sum of squares of $\mathbf{X}$ is higher than that of $\mathbf{y}$ (see expression 7). On the basis of Heij, Groenen and Van Dijk [11] and Gurden [8], we can conclude that α values that lie between .90 and 1 offer the best prediction of future data. However, these studies also showed that a number of data characteristics have an influence on which α value optimizes prediction and recovery, such as the amount of error on the predictors and the criterion, and the number of predictor variables.

2.3. Conclusions and hypotheses

On the basis of the stochastic approach we conclude that in most cases the α value resulting in the maximum likelihood estimates is rather high, unless J is small and $\sigma^2_{\mathbf{E_X}}$ is much larger than $\sigma^2_{\mathbf{e_y}}$. The simulation approach yields similar results in that α values between .90 and 1 lead to the best recovery of the underlying parameter sets and the best prediction of future data. Moreover, both approaches suggest that which α value optimizes a particular performance criterion, depends on the amount of error in $\mathbf{X}$ and $\mathbf{y}$, and on the number of predictors.

It is, however, not only important to know which is the optimal α value given specific data characteristics and for a particular performance measure, but also, how the performance of the model fluctuates across different α values (i.e., how important the selection of α actually is). This question has not been directly addressed before, but the results from Heij, Groenen and Van Dijk [11] and the graphs from Gurden [8] suggest that the importance of α also depends on data characteristics. From a theoretical point of view, it can be expected that the α value will have the most impact when the amount of $\mathbf{X}$ variance explained by each component and their importance for predicting $\mathbf{y}$ differs strongly. Specifically, we predict that in case data contain stronger and weaker components [12,13], PCovR performance will fluctuate more across different α values. In case the weaker components are very relevant for predicting $\mathbf{y}$, the method might fail to recover these weaker components if α is set too high. On the other hand, if the weaker components are not so relevant for predicting $\mathbf{y}$, PCovR might not reveal these components in favor of a noise component, if α is set too low.

3. Simulation study

3.1. Design and procedure

To test the hypotheses that were mentioned in the previous section, we conducted a simulation study in which the following factors were manipulated and fully crossed between datasets (with number of levels between parentheses):

1. The number of observations (2): 50, 100
2. The number of predictors (2): 24, 48
3. The number of underlying components (2): 2, 4

[1] These regression weights can be derived from the PCovR output by multiplying the weight matrix with the regression weights of the components, but they are less insightful when there are many predictor variables.

4. Proportion of error on **X** (3): 5%, 25%, 45%
5. Proportion of error on **y** (3): 5%, 25%, 45%
6. Proportion of structural variance in **X** accounted for by components that are more vs. less relevant for predicting **y** (5): 8% vs. 92%, 33% vs. 67%, 50% vs. 50%, 67% vs. 33%, 92% vs. 8% (see Table 2). This factor was included to examine the combined effect of strength and relevance of the components.

For each cell of the factorial design, 20 datasets were constructed as follows:

$$\mathbf{X} = \mathbf{T}\mathbf{P}_{\mathbf{X}} + \mathbf{E}_{\mathbf{X}} \tag{11}$$

$$\mathbf{y} = \mathbf{T}\mathbf{p}_{\mathbf{y}} + \mathbf{e}_{\mathbf{y}}, \tag{12}$$

where the component score matrix **T** was generated by randomly sampling from a standard normal distribution. Subsequently, we standardized and orthogonalized the columns of **T**. The loading matrix $\mathbf{P}_{\mathbf{X}}$ was created in such a way that every variable loads perfectly (i.e. loading of one) on one component only and has zero-loadings on all other components, but varying the number of variables per component, as displayed in Table 2 (factors 2, 3 and 6). The regression weights in $\mathbf{p}_{\mathbf{y}}$ were set to the values shown in Table 2. Combining factor 6 and the regression weights results in the manipulation of the strength (i.e., amount of variance explained in **X**) and the relevance (i.e., amount of variance explained in **y**) of the different components. For instance, if four components are simulated, the 8% vs. 92% situation pertains to two weak but highly relevant components (each explaining 4% of the variance of **X**) and two strong but less relevant components (each explaining 46%). We sampled the error matrices $\mathbf{E}_{\mathbf{x}}$ and $\mathbf{e}_{\mathbf{y}}$ from a standard normal distribution and standardized and orthogonalized their columns. Moreover, we ensured that the correlations between the columns of **T** and $\mathbf{e}_{\mathbf{y}}$ equal zero. Subsequently, the error matrices, the loadings, and the regression weights were rescaled to obtain predictor and criterion variables that contain the correct proportion of error variance (factors 4 and 5) when computed across all variables. We allowed for differences in the exact amount of error variance per variable, however, by altering the scaling factor per column of $\mathbf{E}_{\mathbf{x}}$ such that one third of the variables contained less error variance and one third more. Finally, the resulting matrices **X** and **y** were standardized.

The resulting 7200 datasets were analyzed with PCovR, using 100 different α values evenly spread between .01 to 1. After each analysis, we applied a Quartimin rotation [14], which minimizes the criterion

$$f(\mathbf{P}_{\mathbf{x}}) = \sum_{j=1}^{J} \sum_{r_1=1}^{r} \sum_{r_2 \neq r_1}^{r} p_{jr_1}^2 p_{jr_2}^2 \tag{13}$$

and allows for correlated components, to both the estimated and the true loading matrix and counter rotated the corresponding component scores and regression weights accordingly. We opted for Quartimin, because in a pilot study it could recover simple structure loading matrices better than Geomin [15] and Varimax [16].

Next, we computed two performance measures for each dataset and each α value. In order to investigate how well PCovR was able to recover the true components, we computed the mean Tucker congruence coefficient [17] between the true and estimated component scores:

$$phi_{\mathbf{T}} = \frac{\sum_{r=1}^{R} \left| \Phi\left(\mathbf{t}_r^{(T)}, \mathbf{t}_r^{(M)}\right) \right|}{R}, \tag{14}$$

with Φ denoting the Tucker congruence coefficient, $\mathbf{t}_r^{(T)}$ the rth component of the true component score matrix, and $\mathbf{t}_r^{(M)}$ the rth component of the estimated component score matrix.

Second, we examined to which extent PCovR was able to recover the true **y** scores, by computing the following badness-of-recovery measure $bor_{\mathbf{y}}$:

$$bor_{\mathbf{y}} = \frac{\left\| \mathbf{y}^{(M)} - \mathbf{y}^{(T)} \right\|^2}{N}, \tag{15}$$

with $\mathbf{y}^{(M)}$ and $\mathbf{y}^{(T)}$ indicating the estimated and the true criterion, respectively. Note that this measure may also yield insight into the predictive success of PCovR. Indeed, we expect a close relationship between recovery of true scores and prediction of future data, since good recovery of true scores implies that the method is not prone to error fitting.

3.2. Results

First, we checked whether our hypotheses about which α value will lead to the best recovery of the underlying components and to good prediction of future data are correct. Figs. 1 and 2 show the α values that are optimal in that they on average lead to the best recovery of the component scores and true **y** scores (note that the performance measures themselves are not displayed), respectively. We can conclude that the optimal α mainly depends on the number of predictors and on the relative amount of error on **X** and **y**. Specifically, as predicted, the α that optimizes the recovery is lower if **X** contains few predictors and a lot of error in comparison to **y**. Regarding the other manipulated factors, Figs. 1 and 2 suggest that the optimal α value seems to depend somewhat on the weakness and the relevance of the components, but the pattern changes across the conditions and many confidence intervals clearly overlap. Importantly, the obtained optimal α values were in most cases relatively close to the α_{ML} values, computed on the basis of the number of predictor variables and the error variances (indicated by the gray line).

Table 2
Number of high loading variables and percentages of structural **X** variance explained per component (between parentheses) in different strength and relevance conditions with corresponding regression weights $\mathbf{p}_{\mathbf{y}}$.

		$R=2$		$R=4$			
		1	2	1	2	3	4
$J=24$	8% vs. 92%	2 (8%)	22 (92%)	1 (4%)	1 (4%)	11 (46%)	11 (46%)
	33% vs. 67%	8 (33%)	16 (67%)	4 (17%)	4 (17%)	8 (33%)	8 (33%)
	50% vs. 50%	12 (50%)	12 (50%)	6 (25%)	6 (25%)	6 (25%)	6 (25%)
	67% vs. 33%	16 (67%)	8 (33%)	8 (33%)	8 (33%)	4 (17%)	4 (17%)
	92% vs. 8%	22 (92%)	2 (8%)	11 (46%)	11 (46%)	1 (4%)	1 (4%)
$J=48$	8% vs. 92%	4 (8%)	44 (92%)	2 (4%)	2 (4%)	22 (46%)	22 (46%)
	33% vs. 67%	16 (33%)	32 (67%)	8 (17%)	8 (17%)	16 (33%)	16 (33%)
	50% vs. 50%	24 (50%)	24 (50%)	12 (25%)	12 (25%)	12 (25%)	12 (25%)
	67% vs. 33%	32 (67%)	16 (33%)	16 (33%)	16 (33%)	8 (17%)	8 (17%)
	92% vs. 8%	44 (92%)	4 (8%)	22 (46%)	22 (46%)	2 (4%)	2 (4%)
$\mathbf{p}_{\mathbf{y}}$		.90	.18	.69	.62	−.09	−.17

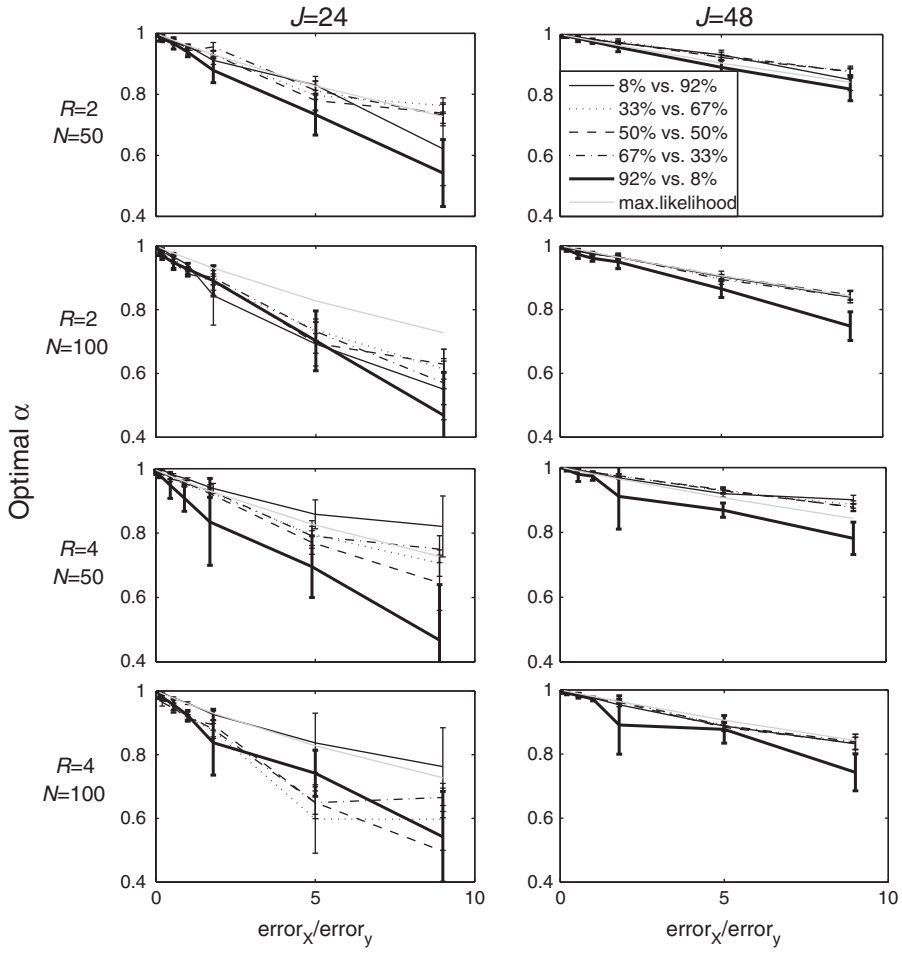


Fig. 1. Overview of α values that yield the highest average recovery of the component scores, in different conditions, with error bars indicating the 95% confidence intervals. The gray line shows the values that can be expected based on Eq. (10).

Second, we investigated the extent to which the recovery of the component scores and the true **y** scores depends on α and how this is influenced by the manipulated data characteristics. To this end, we first conducted an ANOVA with the recovery of the component scores as the dependent variable and the α value as the within-subject factor and the six data characteristics as between-subject factors. Interaction effects for which the partial eta-squared statistic η_p^2 [18] is larger than .25 were the effect between α and the amount of error on the criterion ($\eta_p^2 = .67$), that between α and the variance accounted for by more and less relevant components ($\eta_p^2 = .70$), and that between α and the number of components ($\eta_p^2 = .41$): α has considerable influence when **y** contains more error, and when the number of components is lower (see Fig. 3), which can be understood by realizing that in case of four components the difference between strong and weak components in amount of variance explained becomes smaller. The interaction effect between α and the strength and the relevance of the components implies that the main effect of the strength and relevance of the components – best recovery for 50% vs. 50%, closely followed by the 67% vs. 33% and 33% vs. 67% conditions, and worse recovery for the more difficult 8% vs. 92% and 92% vs. 8% conditions – is more outspoken for smaller and larger α values. Moreover, all associated three-way interactions explain a lot of variance ($.26 \leq \eta_p^2 \leq .54$), in that the effect of α becomes more pronounced when these data characteristics occur together.

We conducted a similar ANOVA for the recovery of the true **y** scores and obtained again strong interaction effects between α and the amount of error on the criterion ($\eta_p^2 = .91$), and between α and the variance accounted for by the different components ($\eta_p^2 = .56$): the more error on the criterion and the more extreme the difference in variance accounted for by the components, the more important the choice of α becomes. Moreover, the interaction effects between α and the number of predictor variables ($\eta_p^2 = .76$) and the number of observations ($\eta_p^2 = .70$), and all associated three-way interactions ($.20 \leq \eta_p^2 \leq .51$) were noteworthy: α has more influence when the dataset becomes more square (i.e., N/J closer to 1), but the $bor_{\mathbf{y}}$ value for the optimal α value is better, when more predictor variables are available.

Comparing the latter results to our hypotheses, we conclude that the prediction that in case some of the components are weaker, the quality of the obtained estimates strongly depends on α, is clearly supported, but also qualified. Indeed, in these cases, somewhat irrespective of the relevance of the components, α should not be too high, but also not too low if **y** contains more error (see Figs. 3 and 4). Moreover, the effect of α seems to depend on the performance measure under consideration: in particular, for the recovery of the component scores, α has only a minor influence when differences in strength are not extreme, as the recovery of the component scores is almost always higher than .95, implying that the true and estimated components can be considered equal [19]. For the recovery of the true **y** scores, however, α always seems to matter if the data contain somewhat more error on the criterion.

It might seem surprising that the worst recovery of component scores is reached for the 92% vs. 8%-condition (Fig. 3), while the worst recovery of true criterion scores is attained for the 8% vs.

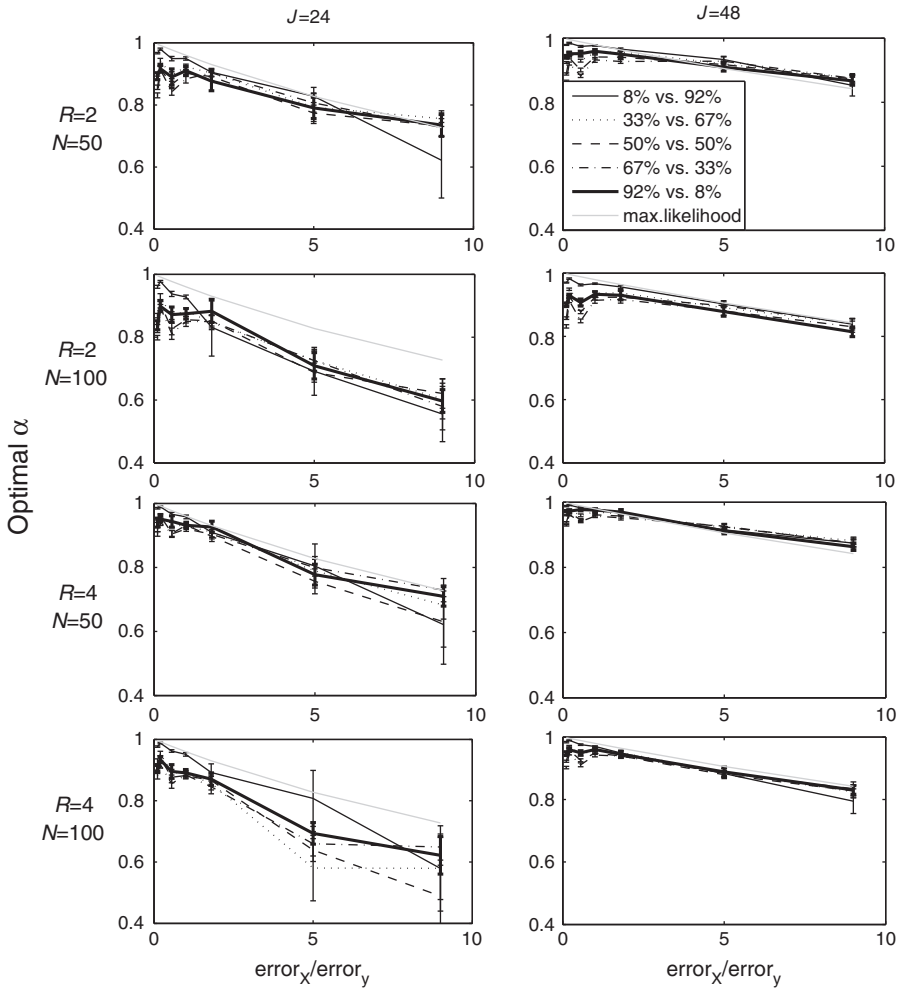


Fig. 2. Overview of α values that yield the highest average recovery of the true criterion scores, in different conditions, with error bars indicating the 95% confidence intervals. The gray line shows the values that can be expected based on Eq. (10).

92%-condition (Fig. 4). This can, however, be explained by the nature of both measures. The recovery of the component scores is evaluated by averaging the obtained Tucker congruence coefficient across the (two or four) components. In the 92% vs. 8%-condition, the strong and relevant components are recovered very well and the weak and less relevant components relatively bad, leading to a relatively low average. On the other hand, in the 8% vs. 92%-condition, all components are recovered reasonably well, because they are either strong or relevant, implying that the average congruence is rather high. The recovery of the true criterion scores, however, is mainly influenced by the most relevant components. As we just argued that these components are recovered better in the 92% vs. 8%-condition than in the 8% vs. 92%-condition, the first condition will lead to better recovery of the true criterion scores than the last.

4. Discussion

In this paper, we discussed and integrated a number of studies which, from different angles, yielded insight into the role of α in PCovR analysis. On the basis of these studies, we predicted that the α value that leads to maximum likelihood estimates, to the best recovery of underlying parameters or the best prediction of future data, depends at least on the number of predictors and on the ratio of the amount of error on the predictors and the amount of error on the criterion. These hypotheses were confirmed by means of an extensive simulation study, in which we examined the recovery of the underlying component scores and the recovery of the underlying criterion scores. Therefore, in case researchers can reliably estimate the amount of error on the predictors and the criterion, they can compute the α_{ML} value for their data by means of expression (10). When no accurate information on the amount of error is available, cross-validation may be the best option for assessing the optimal α value, since good prediction of future data and good recovery of the underlying model parameters usually go together. The latter approach was also advocated by Smilde and Kiers [7] in the context of Multiway Covariates Regression.

While studies up to now focused almost exclusively on which α value is optimal in some sense, an equally or maybe even more important question, as it challenges the key principle behind PCovR, concerns the extent to which different α values really change the obtained results. In this regard, we argued that the importance of α depends on the interplay between the strength of the components (how much variance do they explain) and their relevance for predicting the criterion. Specifically, we predicted that α is of limited importance in case all components are relatively strong. Our simulation results were in line with these conjectures and showed that α has the most influence when the components strongly differ in strength and relevance. However, when recovery of the criterion values is to be optimized, we revealed that the form of the data matrix is important as well as is the amount of error on the criterion variable: when the number of observations approaches the number of variables and the data contain somewhat more error, α plays a key role, even when all components under consideration are relatively strong.

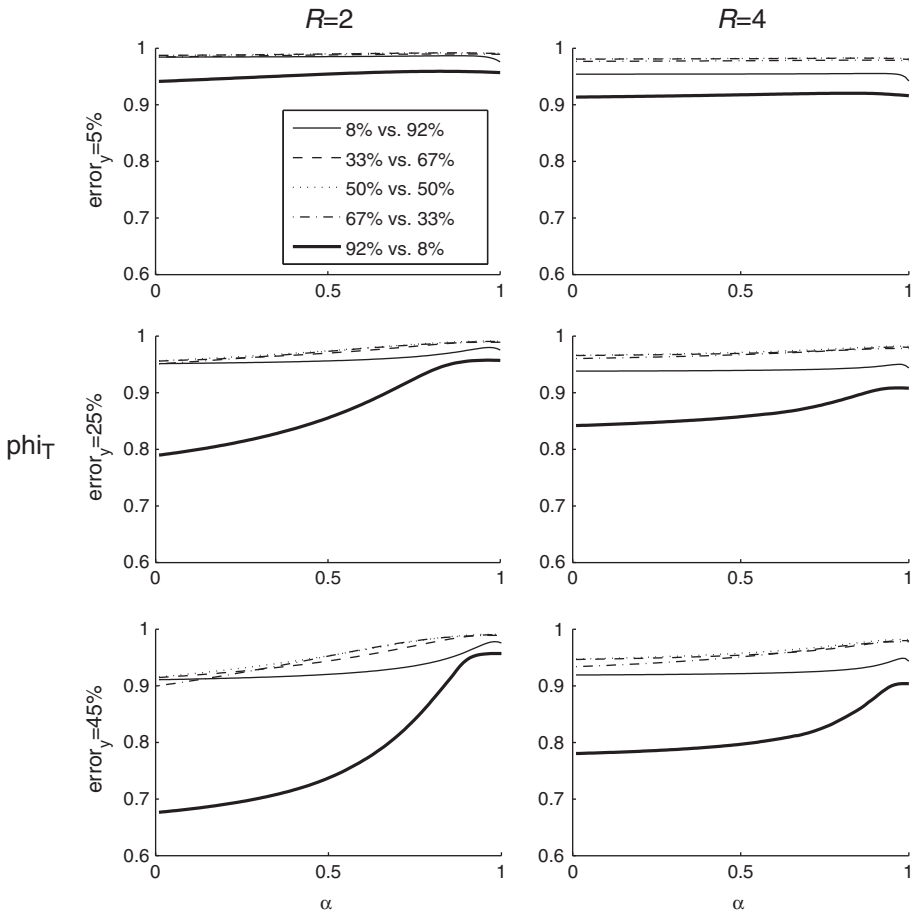


Fig. 3. Average phi value of the component scores ($\mathbf{phi_T}$) in function of α in different conditions for datasets with a fixed regression weight vector.

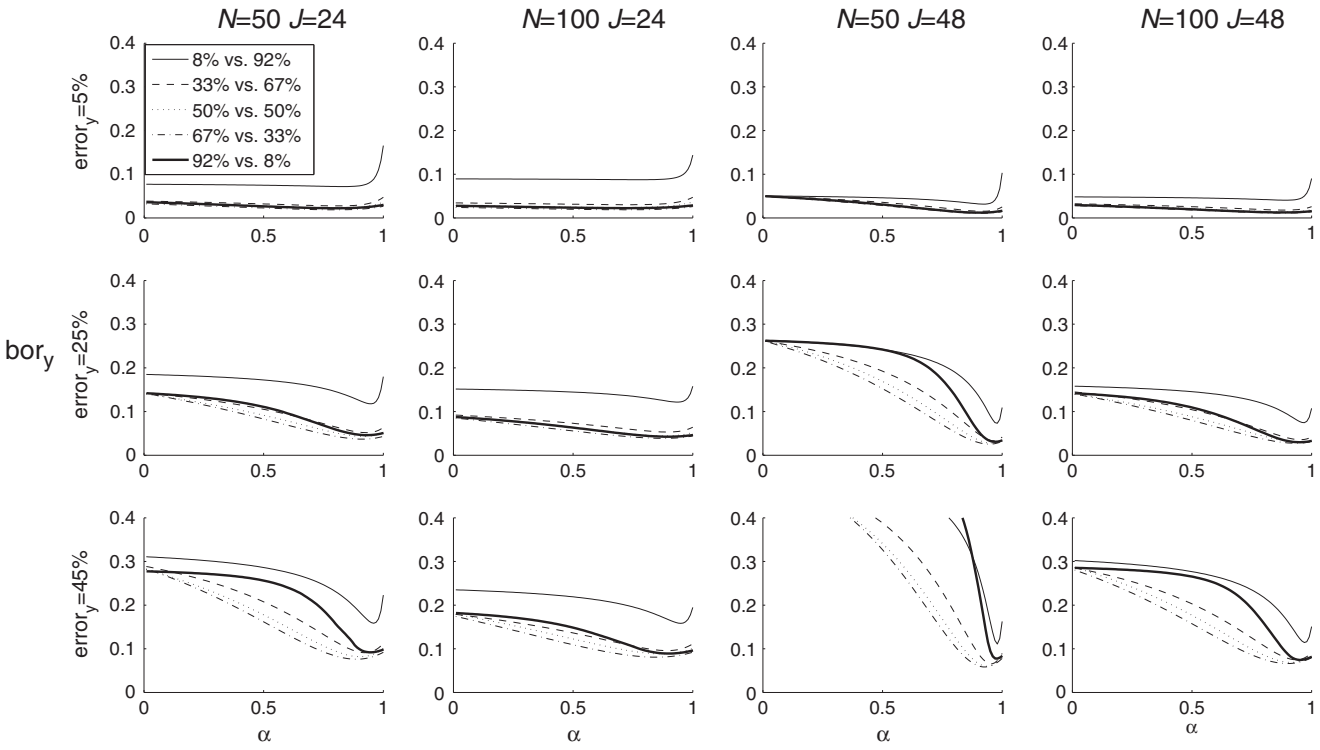


Fig. 4. Average recovery of the true criterion scores ($\mathbf{bor_y}$) in function of α in different conditions for datasets with a fixed regression weight vector.

Acknowledgments

The research reported in this paper was supported by the Research Council of KU Leuven (GOA/2010/02). The simulation study was conducted using the infrastructure of the VSC - Flemish Supercomputer Center, funded by the Hercules foundation and the Flemish Government - department EWI.

References

[1] R. Bro, H.H. Nielsen, G. Stefánsson, T. Skåra, A phenomenological study of ripening of salted herring. Assessing homogeneity of data from different countries and laboratories, Journal of Chemometrics 16 (2002) 81–88.
[2] H.A.L. Kiers, A.K. Smilde, A comparison of various methods for multivariate regression with highly collinear variables, Statistical Methods and Applications 16 (2007) 193–228.
[3] R. Tibshirani, Regression, shrinking and selection via the Lasso, Journal of the Royal Statistical Society: Series B Methodological 58 (1996) 267–288.
[4] M. Aldrin, Reduced-rank regression, in: A.H. El-Shaarawi, W.W. Piegorsch (Eds.), Encyclopedia of Environmetrics, John Wiley & Sons, Chichester, 2002, pp. 1724–1728.
[5] I.T. Jolliffe, A note on the use of principal components in regression, Applied Statistics 31 (1982) 300–303.
[6] S. De Jong, H.A.L. Kiers, Principal covariates regression: part I. Theory, Chemometrics and Intelligent Laboratory Systems 14 (1992) 155–164.
[7] A.K. Smilde, H.A.L. Kiers, Multiway covariates regression models, Journal of Chemometrics 13 (1999) 31–48.
[8] S. Gurden, Multiway covariates regression: discussion facilitator (unpublished).
[9] T.F. Wilderjans, E. Ceulemans, I. Van Mechelen, Simultaneous analysis of coupled data blocks differing in size: a comparison of two weighting schemes, Computational Statistics and Data Analysis 53 (2009) 1086–1098.
[10] T.F. Wilderjans, E. Ceulemans, I. Van Mechelen, R.A. van den Berg, Simultaneous analysis of coupled data matrices subject to different amounts of noise, British Journal of Mathematical and Statistical Psychology 64 (2011) 277–290.
[11] C. Heij, P.J.F. Groenen, D. Van Dijk, Forecast comparison of principal component regression and principal covariate regression, Computational Statistics and Data Analysis 51 (2007) 3612–3625.
[12] E. Ceulemans, H.A.L. Kiers, Discriminating between strong and weak structures in three-mode principal component analysis, British Journal of Mathematical and Statistical Psychology 62 (2009) 601–620.
[13] L.R. Tucker, R.F. Koopman, R.L. Linn, Evaluation of factor analytic research procedures by means of simulated correlation matrices, Psychometrika 34 (1969) 421–459.
[14] J.B. Carroll, An analytical solution for approximating simple structure in factor analysis, Psychometrika 18 (1953) 23–38.
[15] A. Yates, Multivariate exploratory data analysis: a perspective on exploratory factor analysis, SUNY Press, New York, 1987.
[16] H.F. Kaiser, The varimax criterion for analytic rotation in factor analysis, Psychometrika 23 (1958) 187–200.
[17] B. Korth, L.R. Tucker, The distribution of chance congruence coefficients from simulated data, Psychometrika 40 (1975) 361–372.
[18] J. Cohen, Eta-squared and partial eta-squared in fixed factor Anova designs, Educational and Psychological Measurement 33 (1973) 107–112.
[19] U. Lorenzo-Seva, J.M.F. ten Berge, Tucker's congruence coefficient as a meaningful index of factor similarity, Methodology 2 (2006) 57–64.