Some alternatives to PLS

Kiers, Henk A.L.

Published in:
Atti del Convegno Intermedio della Società Italiana di Statistica "Analisi Statistica Multivariata per le Scienze Economico-Sociali, le Scienze Naturali e la Tecnologia"

Publication date:
2003

Citation for published version (APA):
Kiers, H. A. L. (2003). Some alternatives to PLS. In *Atti del Convegno Intermedio della Società Italiana di Statistica "Analisi Statistica Multivariata per le Scienze Economico-Sociali, le Scienze Naturali e la Tecnologia": Napoli 9–11 giugno 2003* (pp. 171-182). Napoli: RCE Edizioni.

Some Alternatives to PLS

Qualche Alternativo di PLS

Henk A.L. Kiers,
Heymans Institute, University of Groningen, Grote Kruisstraat 2/1,
9712 TS Groningen, The Netherlands, e-mail: h.a.l. kiers@ppsw.rug.nl

Riassunto: In questo articolo vengono presentati alcuni metodi per un caso particolare di regressione in cui le variabili indipendenti sono sintetizzate prima di essere utilizzate come predittori. In tale contesto, infatti, l'obiettivo non è semplicemente spiegare le variabili risposta, ma anche sintetizzare le variabili indipendenti. A questo proposito, il metodo maggiormente utilizzato è il PLS, sebbene ne esistano anche altri, come il PcovR. Nostro obiettivo è sviluppare un ulteriore metodo, denominato "Power Regression", che verrà discusso in dettaglio anche in riferimento agli aspetti computazionali.

Keywords: regression, pls, PCovR, Power Regression

1. Introduction

Multiple regression aims at finding an optimal rule for predicting scores on a criterion variable (*dependent variable*) on the basis of scores on a number of predictor variables (*independent variables*). The prediction rule is obtained by analyzing data on a training sample, for which scores on both the predictor variables and the criterion variable are available, and finding that linear combination of variables that approximates most closely the scores on the criterion variable. The regression weights (i.e., the weights used to form the optimal linear combination) then define the prediction rule, which is meant to be useful in situations where it is desired to estimate the *unknown* scores on a criterion variable, while *only* the scores on the predictor variables are available. Such situations are very common, since they arise as soon as a variable of much practical relevance is hard or even impossible to measure, whereas other variables that can be used to predict it are easily available. A common example is the situation where the criterion variable refers to future sales of a product (which by definition cannot be measured now), that could be predicted by known attributes of the product or the prospective buyers. Obviously, the usefulness of a prediction rule does not reside in its performance for the data on the basis of which it was obtained, but in its performance on other (e.g., future) data. In other words, a regression rule should primarily have good generalizability properties.

It is well known that the prediction rule resulting from ordinary multiple regression is rather prone to lack of generalizability, as will be explained in Section 2. For this reason, various alternatives to regression, sometimes called biased regression techniques, have been proposed. The best known of these are PLS (*Partial Least Squares*, e.g., see Wold, 1966, Wold et al. 1984, Martens & Naes, 1989) and PCR (*Principal Component Regression*, see Coxe, 1986, Martens & Naes, 1989), but several other methods have also been developed (e.g., ILS by Frank, 1987; Continuum

Regression by Stone & Brooks, 1990; The Curds and Whey procedure by Breiman & Friedman, 1997, and various recent, more specialized procedures, e.g. see Esposito Vinzi et al, 2001). As will also be explained in Section 2, an important improvement of the performance of the prediction rule can be expected if it is based on a method that not only optimizes variance explained in the criterion variable(s), but also in the predictor variables. This idea seems to underly the success of PLS, even though in PLS this is not *explicitly*. Here, I will focus on alternatives to PLS that explicitly optimize a compromise between explained variance in the predictor variables and explained variance in the criterion variables. The first such method is PCovR (*Principal Covariates Regression,* De Jong & Kiers, 1992). A second method is a new one (although already anticipated by De Jong & Kiers, 1992, pp. 160-161), which will be called *Power Regression*. This combines the principles underlying PCovR in a different, possibly more robust way. First, however, a comparison of PCovR to some other techniques, among which PLS will be given in Section 3. Power Regression is introduced in Section 4, and it is indicated what will be the relative advantages and disadvantages of this method. Section 5 is devoted to some three-way extensions of biased regression techniques. Finally, the paper is finished with a conclusion in Section 6.

2. How to avoid ungeneralizable prediction rules

As mentioned above, a prediction rule should not only perform well for the data on the basis of which it was obtained (the “training data”), but also for other data. In practice, however, it frequency happens that the prediction rule works well only for the training data, and not for other data. This happens when the number of predictor variables is large relative to the number of observation units, and the predictor variables are correlated. In such cases, the sheer multitude of predictor variables will often guarantee that regression weights can be found that lead to good prediction rules, simply because the large number of vectors with scores on predictor variables span a high dimensional space that is likely to capture most of the information in the vector of scores on the criterion variable(s). Indeed, when there are at least as many predictor variables as observation units, then it is practically guaranteed that regression weights can be found that lead to a prediction rule that ‘predicts’ the criterion scores in the training data perfectly. This clarifies that regression is a greedy technique, capitalizing on peculiarities in the training data, that need, however, not hold for any other data. This problem of regression is aggravated when the correlation between the predictor variables is high, because then, in principle, using several predictor variables cannot be expected to improve the prediction of a criterion variable much over using only of them, because they all convey largely the same information. However, the regression technique finds regression weights that work optimally for the training data, and for this purpose, it then exploits and inflates minimal differences between the predictor variables to get optimal predictions of the criterion variable(s). This leads to what is also called the ‘bouncing beta’ problem, where regression weights become very large due to multicollinearity of the predictor variables.

The bouncing beta problem received much attention in the literature, and has been approached directly by techniques that explicitly aim at ‘shrinkage’ of the regression weights. An example is ridge regression (Hoerl & Kennard, 1970), which can easily be seen as a method that fits the regression model in the least squares sense with a penalty on the size of the regression weights. However, a mere shrinkage by itself may not solve the problem of poor performance in data other than the training data. Obviously, no guaranteed solution can be given to this problem, because the relation between predictor variables and criterion scores in such other data is not known. However, considering that the use of many predictor variables may easily lead to capitalizing on peculiarities in the training data, an approach that could help here is to reduce the number of predictor variables by combining predictor variables that convey more or less the same information, into summarizers. According to the above idea, techniques have been developed that aim at optimal prediction of the criterion variable(s) by means of preliminarily or simultaneously found summarizers of the predictor variables. The quality of the prediction of the criterion variable(s) (usually denoted as *Y*) is often expressed as the proportion of explained variance R_Y^2, while the quality of the summary of the predictor variables (denoted by *X*) is often expressed as the proportion of explained variance R_X^2. The present paper focuses on techniques that implicitly or explicitly aim to maximize a compromise of R_X^2 and R_Y^2.

3. PCR, PLS, PcovR and related techniques

3.1. Notation

The training data set consists of scores on predictor variables and one or more criterion variables. The notation used here is:
$\mathbf{X}$ = $n{\times}p$ matrix $\mathbf{X}$ with scores of *n* observation units on *p* predictor variables,
$\mathbf{Y}$ = $n{\times}m$ matrix with scores of *n* observation units on *m* criterion variables,
$\mathbf{y}$ = $n{\times}1$ vector with scores on the criterion variable (in case there is only one).
It should be noted that $\mathbf{X}$ and $\mathbf{Y}$ or $\mathbf{y}$ are assumed to be centered columnwise.
In the various regression methods we will encounter the following matrices:
$\mathbf{T}$ = $n{\times}r$ matrix with scores of *n* observation units on *r* summarizers, with $\mathbf{T} = \mathbf{XW}$,
$\mathbf{W}$ = $p{\times}r$ matrix with component weights,
$\mathbf{B}$ ($\mathbf{b}$) = $p{\times}m$ ($p{\times}1$) matrix of weights for regression of $\mathbf{Y}$ (or $\mathbf{y}$) on $\mathbf{X}$,
$\mathbf{P_Y}$ ($\mathbf{p_y}$) = $r{\times}m$ ($r{\times}1$) matrix of weights for regression of $\mathbf{Y}$ on $\mathbf{T}$,
$\mathbf{P_X}$ = $r{\times}p$ matrix of weights for regression of $\mathbf{X}$ on $\mathbf{T}$.

3.2. PCR

The most straightforward implementation of the idea of finding good summarizers, and finding a good prediction of criterion scores on the basis of these summarizers is PCR. In PCR the predictor variables are summarized by means of a limited number of principal components (with component weights collected in $\mathbf{W}$, and component scores in $\mathbf{T}$), and next these principal components are used as predictors in a regression of the criterion variable(s) on the principal components, by minimizing $\|\mathbf{Y}-\mathbf{TP_Y}\|^2$ = $\|\mathbf{Y}-\mathbf{XWP_Y}\|^2$. Thus, the regression weights resulting from PCR are given by $\mathbf{B} = \mathbf{WP_Y}$, and these are the weights that define the prediction rules to be used with other data as well. Because PCR uses principal components, the components $\mathbf{T}$ maximize R_X^2 =

$1-\|\mathbf{X}-\mathbf{TP_X}\|^2/\|\mathbf{X}\|^2$, where $\mathbf{P_X} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{X}$. Given the components in **T**, next $R_Y^2 = 1-\|\mathbf{Y}-\mathbf{TP_Y}\|^2/\|\mathbf{Y}\|^2$ is maximized over the regression weights $\mathbf{P_Y}$, hence $\mathbf{P_Y} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{Y}$.

3.3. PLS

PLS aims at simultaneously, rather than sequentially, summarizing the predictor variables and performing regression on the summarizers. In case of more than one criterion variable, PLS also summarizers these criterion variables. It finds in total *r* pairs of summarizers (each pair consisting of one summarizer for **X** and one for **Y**), and finds each pair by a separate iterative procedure. For the first pair of summarizers, this procedure is as follows:

```
Step 0.Initialize a weights vector $\mathbf{u}$
Step 1. $\mathbf{w} = \mathbf{X}'\mathbf{u} / \| \mathbf{X}'\mathbf{u} \|$
Step 2. $\mathbf{t} = \mathbf{Xw}$
Step 3. $\mathbf{q} = \mathbf{Y}'\mathbf{t} / \| \mathbf{Y}'\mathbf{t} \|$
Step 4. $\mathbf{u} = \mathbf{Yq}$.
Repeat Steps 1 through 4, until convergence.
```

In case **Y** has only one column, Steps 3 and 4 can be dropped, and **u** can be set equal to **y**. It can be shown (e.g., see Manne, 1987) that the solution for vectors **w** and **q** (which are both normalized to unit length) maximizes $\mathbf{w}'\mathbf{X}'\mathbf{Yq} = \mathbf{t}'\mathbf{u}$ over all unit length vectors **w** and **q**, hence subject to this constraint, the first PLS summarizers maximize the covariance between them, and hence also its square. In case of a single criterion variable (**y**), it hence maximizes $\text{cov}^2(\mathbf{y},\mathbf{t}) = \text{cor}^2(\mathbf{y},\mathbf{t})\times\text{var}(\mathbf{t}) = \text{cor}^2(\mathbf{y},\mathbf{Xw})\times\text{var}(\mathbf{Xw}) = R_y^2\times\text{var}(\mathbf{Xw})$. Hence, the first PLS solution maximizes the product of the proportion of explained variance in **y** and the variance of the summarizer of **X**. Relatively high variance of **Xw** may often accompany a high R_X^2 (proportion of explained variance of **X**), but does not necessarily do so, hence the PLS criterion only *indirectly* aims at optimizing explained variance of **X**.

Subsequent pairs of summarizers are obtained after first deflating the data matrices **X** and **Y** by subtracting the predictions on the basis of the previous summarizers of the predictor variables, and then applying the same procedure to these deflated matrices. An alternative to this approach for finding subsequent summarizers is offered by the SIMPLS algorithm (De Jong, 1993), which avoids the deflations and replaces these by a procedure that ensures the summarizers of the predictor variables to be mutually uncorrelated, which leads to slightly different solutions than PLS gives.

3.4. PCOVR

As has been seen above, PCR uses principal components of **X** and hence the components **T** maximize $R_X^2 = 1-\|\mathbf{X}-\mathbf{TP_X}\|^2/\|\mathbf{X}\|^2$. These predictor variables were summarized in an attempt to reduce the number of predictor variables and to avoid finding regression weights that overexploit subtle differences in the scores on the predictor variables in the training data. However, by reducing the information in the predictor variables, one might possibly have thrown away also some information in the predictor variables that is (also outside the training data set) relevant in the prediction of criterion scores. Therefore, rather than finding summarizers (‘components’) that focus on maximizing R_X^2 (as in PCR), PcovR aims at finding components that not only summarize the scores on the predictor variables well, but also predict the scores on the criterion variables well. Specifically, PcovR maximizes $\alpha R_X^2+(1-\alpha)R_Y^2$, where α is a parameter that has to be chosen by the user. Equivalently, it can be said that PcovR minimizes

$$f(\mathbf{W}) = \alpha \parallel \mathbf{X} - \mathbf{TP_X} \parallel^2 + (1-\alpha) \parallel \mathbf{Y} - \mathbf{TP_Y} \parallel^2$$

$$= \alpha \parallel \mathbf{X} - \mathbf{XWP_X} \parallel^2 + (1-\alpha) \parallel \mathbf{Y} - \mathbf{XWP_Y} \parallel^2, \qquad (1)$$

where $\mathbf{P_X} = (\mathbf{W'X'XW})^{-1}\mathbf{W'X'X}$ and $\mathbf{P_Y} = (\mathbf{W'X'XW})^{-1}\mathbf{W'X'Y}$. Clearly, when α=1, we focus entirely on summarizing **X**, which gives us PCR. When α=0, we focus entirely on predicting **Y**, which gives us reduced rank regression, or, when the number of components is at least as large as the number of criterion variables, ordinary regression. The more interesting cases are those where α is in between the extremes, because then the method finds components that aims at finding components **T** that simultaneously summarize **X** well and predict **Y** well.

The solution for the minimization of (1) can be obtained noniteratively. For identification, the nonrestrictive constraint $\mathbf{T'T}=\mathbf{W'X'XW}=\mathbf{I}$ is imposed. Let $\mathbf{H_X}$ denote the projector $\mathbf{X}(\mathbf{X'X})^+\mathbf{X}$, and let **E** contain the first *r* eigenvectors of

$$\mathbf{G} = \alpha\, \mathbf{XX'} + (1-\alpha)\, \mathbf{H_X YY' H_X}, \qquad (2)$$

then the solution is given by $\mathbf{W}=\mathbf{X}^+\mathbf{E}$, and hence $\mathbf{T} = \mathbf{E}$; here the superscript $^+$ denotes the Moore-Penrose inverse. Next $\mathbf{P_Y}$ can be computed as $\mathbf{P_Y} = \mathbf{T'Y}$. The regression weights matrix **B** then is $\mathbf{B} = \mathbf{WP_Y} = \mathbf{X}^+\mathbf{EE'Y}$.

The main problem with PcovR is that it is difficult to make a choice for α. De Jong and Kiers (1992) suggested to use cross-validation to choose α, but practical experience with this approach has not met with unequivocal success. One reason for this may be that the solution depends on the size of the data values. That is, an overall scaling of matrix **X** will change the solution by more than a proportional change of the regression weights, because it affects the contents of the matrix **G** in (2) in a nonproportional way. Actually, rescaling the data, has the same effect as using a different value of α. Hence, the scale of the data, and the choice of α are confounded, and this choice of α should be made after one has decided on a proper way of scaling the data. The problem then is that it is not a priori clear what should be a proper scale of the data, and finding a good value of α becomes somewhat cumbersome. This problem does not occur with methods that are ‘scale free’, that is for which the results are not affected by rescaling the data (except for trivial rescalings of the regression weights). Ordinary regression, PCR and PLS are all scale free methods, but none of them simultaneously optimizes R_X^2 and R_Y^2. The method to be described in the next section is also scale free, and *does* simultaneously optimize R_X^2 and R_Y^2.

4. Power Regression

In Section 3.3 it has been mentioned that PLS finds components that (sequentially) maximize the product of the explained variance of **y** and the variance of the component itself. If one searches components that account well for the information in both **X** and **y** (or, if more than one criterion variable is available, **Y**), the above criterion seems somewhat inconsistent: As mentioned, the criterion involves explained variance of **Y**, whereas it uses the *variance* of **Xw** (rather than the variance of **X** *explained* by **Xw**). In an attempt to find components that both explain well the variance of **X** and that of **Y**, PCovR has been proposed. PCovR maximizes a weighted sum of R_{X}^2 and R_y^2. As mentioned above in Section 3.4, a problem in the application of PCovR is to choose the weights in this weighted sum. In the present section, a different approach to maximizing a compromise of R_X^2 and R_Y^2 is offered (extending on a proposal by De Jong & Kiers, 1992, pp.160-162). This is based on maximizing the *product* of R_X^2 and R_Y^2 (or a sum of such products).

The rationale for maximizing the product of R_X^2 and R_Y^2 rather than the sum is that this will imply that both R_X^2 and R_Y^2 will be reasonably high. This is because one small value can strongly decrease the value of the product criterion, whereas in PcovR, maximizing the sum of R_X^2 and R_Y^2, a small value of the one can more easily be compensated by a high value of the other. Because of this, it can be expected that, when taking the product of R_X^2 and R_Y^2, no weightings of R_X^2 and R_Y^2 are needed to avoid that either R_X^2 or R_Y^2 becomes too small. Moreover, because this method employs only the products of R^2 values, which do not depend on the scale of the data, the method is scale free, and, again, no weightings of R_X^2 and R_Y^2 are needed to compensate for scale differences that unduly affect the solution. This is not to say that it is impossible to implement such weights: Indeed, in a similar way as in continuum regression (Stone & Brooks, 1990), we could take different *powers* of the terms in the product. So a general formulation of the criterion could be (see De Jong & Kiers, 1992) to maximize $(R_X^2)^{2\alpha}(R_Y^2)^{2-\alpha}$, over component weights **W** and hence components **T**. However, here we will only deal with the case where α=½, hence maximizing the product $R_X^2R_Y^2$, because we expect that this will suffice for most practical purposes.

Using a product criterion actually follows the implicit rationale behind PLS. It differs, however, from PLS in that the product involves *explained* variances for **Y** *and* **X**. In this way, we combine the good features of PLS (using the product criterion) and PCovR (using explained variances), and thus hope to get the best of both worlds. We dubbed the method *Power Regression* because of its power to explain variance in both **Y** *and* **X**.

As mentioned above, this approach was first proposed by De Jong and Kiers (1992), but then further ignored, maybe because their algorithm for maximizing this product was somewhat ad hoc, and not known to converge. Moreover, they proposed only an approach where R_X^2 and R_Y^2 are based on all components simultaneously, whereas here also a criterion will be considered where the sum of the products of R_X^2 and R_Y^2 computed per component will be maximized. This can be expected to better avoid solutions in which some components contribute very little to explaining variance in either **X** or **Y**.

4.1. Power Regression using one component

We first describe the Power Regression criterion when only one component is used. The variances of **X** and **Y** explained by a single component **Xw** are given by

$$\|\mathbf{X}\|^2 - \min_{\mathbf{P}} \|\mathbf{X} - \mathbf{XwP}\|^2 = \frac{\mathbf{w}'(\mathbf{X}'\mathbf{X})^2\mathbf{w}}{\mathbf{w}'\mathbf{X}'\mathbf{Xw}} \tag{3}$$

and

$$\|\mathbf{Y}\|^2 - \min_{\mathbf{b}} \|\mathbf{Y} - \mathbf{Xwb}'\|^2 = \frac{\mathbf{w}'\mathbf{X}'\mathbf{YY}'\mathbf{Xw}}{\mathbf{w}'\mathbf{X}'\mathbf{Xw}}, \tag{4}$$

respectively. The proportions of explained variance R_X^2 and R_Y^2 are obtained by division of (3) and (4) by, respectively, $||\mathbf{X}||^2$ and $||\mathbf{Y}||^2$. Then Power Regression using only one component is defined by the regression of **y** on **Xw**, where **w** is found by maximizing

$$R_X^2(\mathbf{w})R_Y^2(\mathbf{w}) = \frac{1}{\|\mathbf{X}\|^2\|\mathbf{Y}\|^2} \frac{\mathbf{w}'(\mathbf{X}'\mathbf{X})^2\mathbf{w}}{\mathbf{w}'\mathbf{X}'\mathbf{Xw}} \frac{\mathbf{w}'\mathbf{X}'\mathbf{YY}'\mathbf{Xw}}{\mathbf{w}'\mathbf{X}'\mathbf{Xw}} \tag{5}$$

over **w**. It should be noted that (5) does not depend on the sum of squares of **w**, hence, we do not have to impose a constraint like **w′w** =1 (as done in PLS) to obtain a sensible solution. As a consequence, we can *arbitrarily* fix the scale in whichever way we like, and, if desired, choose a different scaling after the solution has been found. For convenience, we choose **w′X′Xw**=1, which simplifies the criterion into

$$\mathrm{p}(\mathbf{w}) = (\mathbf{w}'(\mathbf{X}'\mathbf{X})^2\mathbf{w})(\mathbf{w}'\mathbf{X}'\mathbf{YY}'\mathbf{Xw}). \tag{6}$$

To maximize (6) over **w**, subject to **w′X′Xw**=1, we consider two cases:

Case a. The columns of **X** do span the full $\Re^p$ (which is usually the case when $n \leq p$)
Case b. The columns of **X** do not span the full $\Re^p$ (e.g., when $n>p$)

Case a. When **X** spans the full $\Re^p$, any vector **t** can be written as **Xw**, hence, we can replace **Xw** by an arbitrary vector **t**. Then criterion (6) can be written as

$$\mathrm{p}(\mathbf{t}) = (\mathbf{t}'\mathbf{XX}'\mathbf{t})(\mathbf{t}'\mathbf{YY}'\mathbf{t}), \tag{7}$$

which is to be maximized over arbitrary **t**, subject to **t′t**=**w′X′Xw=**1. An algorithm for this maximization problem is described below. Once the optimal **t** is found, we can obtain the optimal **w** as $\mathbf{w}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{t}$.

Case b. When **X** does not span the full $\Re^p$, we define $\tilde{\mathbf{t}} = (\mathbf{X}'\mathbf{X})^{1/2}\mathbf{w}$. It should be noted that any vector $\tilde{\mathbf{t}}$ can be written as $(\mathbf{X}'\mathbf{X})^{1/2}\mathbf{w}$ for a certain vector **w** provided that the inverse of $(\mathbf{X}'\mathbf{X})^{1/2}$ exists, namely by choosing $\mathbf{w}=(\mathbf{X}'\mathbf{X})^{-1/2}\tilde{\mathbf{t}}$. If the inverse does not exist, we can use the Moore-Penrose inverse and we have $\mathbf{w}=((\mathbf{X}'\mathbf{X})^{1/2})^{+}\tilde{\mathbf{t}}$, because, as will be seen later, the optimal $\tilde{\mathbf{t}}$ is in the column space of $(\mathbf{X}'\mathbf{X})^{1/2}$, hence indeed $(\mathbf{X}'\mathbf{X})^{1/2}((\mathbf{X}'\mathbf{X})^{1/2})^{+}\tilde{\mathbf{t}} = \tilde{\mathbf{t}}$; therefore, we will use the Moore-Penrose inverse generally instead of the ordinary inverse. With the above definition of $\tilde{\mathbf{t}}$, the criterion (6) reduces to

$$\begin{aligned} \mathrm{p}(\tilde{\mathbf{t}}) &= (\tilde{\mathbf{t}}'((\mathbf{X}'\mathbf{X})^{1/2})^{+}(\mathbf{X}'\mathbf{X})^{2}((\mathbf{X}'\mathbf{X})^{1/2})^{+}\tilde{\mathbf{t}})(\tilde{\mathbf{t}}'((\mathbf{X}'\mathbf{X})^{1/2})^{+}\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}((\mathbf{X}'\mathbf{X})^{1/2})^{+}\tilde{\mathbf{t}}) \\ &= (\tilde{\mathbf{t}}'\mathbf{X}'\mathbf{X}\tilde{\mathbf{t}})(\tilde{\mathbf{t}}'((\mathbf{X}'\mathbf{X})^{1/2})^{+}\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}((\mathbf{X}'\mathbf{X})^{1/2})^{+}\tilde{\mathbf{t}}), \qquad (8) \end{aligned}$$

which is to be maximized over arbitrary $\tilde{\mathbf{t}}$, subject to $\tilde{\mathbf{t}}'\tilde{\mathbf{t}} = \mathbf{w}'\mathbf{X}'\mathbf{X}\mathbf{w} = 1$. An algorithm for this maximization problem is described below.

Clearly, the optimization problems in (7) and (8) have the same shape, and can be written generally as the maximization of

$$\mathrm{h}(\mathbf{u}) = \mathbf{u}'\mathbf{S}\mathbf{u}\mathbf{u}'\mathbf{T}\mathbf{u}, \qquad (9)$$

subject to $\mathbf{u}'\mathbf{u}=1$, where $\mathbf{S}$ and $\mathbf{T}$ are both positive semi-definite matrices. Upon substituting the appropriate vectors for $\mathbf{u}$ and matrices for $\mathbf{S}$ and $\mathbf{T}$, we reobtain (7) and (8), and it can be seen that in both cases the matrices to be substituted for $\mathbf{S}$ and $\mathbf{T}$ are positive semi-definite. To maximize (9), we use the following iterative algorithm:

Step 1. Initialize $\mathbf{u}^{i}$ (i=0) (e.g., as random vector with unit sum of squares)
Step 2. Compute $\mathrm{h}(\mathbf{u}^{i})$
Step 3. Compute $\mathbf{u}^{i+1}$ as the first eigenvector of $\mathbf{S}\mathbf{u}^{i}\mathbf{u}^{i\prime}\mathbf{T}+\mathbf{T}\mathbf{u}^{i}\mathbf{u}^{i\prime}\mathbf{S}$
Step 4. Compute $\mathrm{h}(\mathbf{u}^{i+1})$; if $\mathrm{h}(\mathbf{u}^{i+1}) - \mathrm{h}(\mathbf{u}^{i}) > \varepsilon\mathrm{h}(\mathbf{u}^{i})$ for some prespecified small value ε (e.g., $\varepsilon=10^{-6}$), then go to Step 3; else consider the algorithm converged.

The above algorithm increases $\mathrm{h}(\mathbf{u})$ monotonically, and because $\mathrm{h}(\mathbf{u}^{i})$ is bounded, the algorithm converges to a stable function value. A proof for the monotonicity of the algorithm (i.e., the fact that $\mathrm{h}(\mathbf{u}^{i+1}) \geq \mathrm{h}(\mathbf{u}^{i})$) is given in the Appendix. It should be noted that the above algorithm can be used for maximizing (7) and (8). If it is used for maximizing (8), at any stage of the algorithm, $\tilde{\mathbf{t}}$ is an eigenvector of

$$\begin{aligned} &\mathbf{X}'\mathbf{X}\tilde{\mathbf{t}}\tilde{\mathbf{t}}'((\mathbf{X}'\mathbf{X})^{1/2})^{+}\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}((\mathbf{X}'\mathbf{X})^{1/2})^{+} \\ &\qquad + ((\mathbf{X}'\mathbf{X})^{1/2})^{+}\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}((\mathbf{X}'\mathbf{X})^{1/2})^{+}\tilde{\mathbf{t}}\tilde{\mathbf{t}}'\mathbf{X}'\mathbf{X}, \qquad (10) \end{aligned}$$

the column space of which is a subspace of that of spanned by $((\mathbf{X}'\mathbf{X})^{1/2})^{+}$, hence $\tilde{\mathbf{t}}$ will always lie in this column space, as was required (see above).

Using the above algorithm, we solve for $\mathbf{t}$ or $\tilde{\mathbf{t}}$, and can next obtain $\mathbf{w}$, and the component scores in $\mathbf{t}=\mathbf{X}\mathbf{w}$. Next, the criterion scores in $\mathbf{Y}$ are regressed onto these component scores, which gives the vector with weights $\mathbf{P}_{\mathbf{Y}}$. Finally, the regression weights matrix $\mathbf{B}$ is computed as $\mathbf{B} = \mathbf{W}\mathbf{P}_{\mathbf{Y}}$, thus yielding the prediction rule for estimating the scores on $\mathbf{Y}$ from those on $\mathbf{X}$.

4.2. Using more than one component

The above method can be generalized to situations where r (r>1) components are used in various ways. Here two approaches are considered. The first is the one proposed by De Jong and Kiers (1992), which consist of maximizing $R_X^2(\mathbf{W})R_Y^2(\mathbf{W})$, where $R_X^2(\mathbf{W})$ and $R_Y^2(\mathbf{W})$ denote the proportions of explained variance using all components jointly. De Jong and Kiers mentioned a simple ad hoc algorithm for maximizing this, but, as has been found in simulations now, this algorithm does not necessarily converge. However, a minor adjustment to this algorithm can be constructed that does lead to a monotonically convergent algorithm. This modification is obtained via a relatively complex derivation based on the majorization approach described by Kiers (1990), which for reasons of space is not given here. The algorithm itself, however, is still fairly simple.

The second approach to handle more than one component consists of maximizing

$$\sum_{l=1}^{r} R_X^2(\mathbf{w}_l)R_Y^2(\mathbf{w}_l) = \text{constant}\times\sum_{l=1}^{r}\frac{\mathbf{w}_l'(\mathbf{X}'\mathbf{X})^2\mathbf{w}_l}{\mathbf{w}_l'\mathbf{X}'\mathbf{X}\mathbf{w}_l}\,\frac{\mathbf{w}_l'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w}_l}{\mathbf{w}_l'\mathbf{X}'\mathbf{X}\mathbf{w}_l} \qquad (11)$$

over $\mathbf{w}_l$, $l=1,...,r$, subject to the constraint $\mathbf{w}_l'\mathbf{X}'\mathbf{X}\mathbf{w}_{l'}$=0 if $l'\neq l$. Thus, the components $\mathbf{X}\mathbf{w}_1$,..., $\mathbf{X}\mathbf{w}_r$ are mutually orthogonal. For convenience, again we scale them such that $\mathbf{w}_l'\mathbf{X}'\mathbf{X}\mathbf{w}_l$=1, $l=1,...,r$. In this criterion we chose to maximize the sum of the products of componentwise R^2-values. This alternative to the approach of maximizing $R_X^2(\mathbf{W})R_Y^2(\mathbf{W})$ is suggested here, because in the latter criterion nothing prevents that some components have hardly any explanatory power in **X** or **Y**, which may in fact lead to poor components after all. In the present approach, the optimality of the sum of products should ensure that *all* components account reasonably well for both **X** and **Y** variance.

To maximize (11) over $\mathbf{w}_1$,..., $\mathbf{w}_r$, we again consider the two cases:

Case a. The columns of **X** do span the full $\Re^p$ (which is usually the case when $n\leq p$)
Case b. The columns of **X** do not span the full $\Re^p$ (e.g., when $n>p$)

<u>Case a</u>. When **X** spans the full $\Re^p$, we define $\mathbf{t}_l\equiv\mathbf{X}\mathbf{w}_l$, with $\mathbf{t}_l'\mathbf{t}_{l'}$=0 and $\mathbf{t}_l'\mathbf{t}_l$=1, and we have to maximize

$$\text{p}(\mathbf{t}_1,..., \mathbf{t}_r)= \sum_{l=1}^{r}\mathbf{t}_l'\mathbf{X}\mathbf{X}'\mathbf{t}_l\mathbf{t}_l'\mathbf{Y}\mathbf{Y}'\mathbf{t}_l\ . \qquad (12)$$

After the optimization of $\mathbf{t}_1$,...,$\mathbf{t}_r$, we can obtain $\mathbf{w}_1$,...,$\mathbf{w}_r$ as $\mathbf{w}_l=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{t}_l$, l=1,...,r

<u>Case b</u>. When **X** does not span the full $\Re^p$, we define $\tilde{\mathbf{t}}_l=(\mathbf{X}'\mathbf{X})^{1/2}\mathbf{w}_l$, l=1,...,r, with $\tilde{\mathbf{t}}_l'\tilde{\mathbf{t}}_{l'}$=0 and $\tilde{\mathbf{t}}_l'\tilde{\mathbf{t}}_l$=1, and we have to maximize

$$\mathrm{p}(\tilde{\mathbf{t}}_1,\ldots,\tilde{\mathbf{t}}_r) = \sum_{l=1}^{r}\tilde{\mathbf{t}}_l'\mathbf{X}'\mathbf{X}\tilde{\mathbf{t}}_l\tilde{\mathbf{t}}_l'((\mathbf{X}'\mathbf{X})^{1/2})^{+}\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}((\mathbf{X}'\mathbf{X})^{1/2})^{+}\tilde{\mathbf{t}}_l\ . \qquad (13)$$

In both cases we have to optimize a function of the form

$$\mathrm{h}(\mathbf{U}) = \sum_{l=1}^{r}\mathbf{u}_l'\mathbf{S}\mathbf{u}_l\mathbf{u}_l'\mathbf{T}\mathbf{u}_l\ , \qquad (14)$$

subject to $\mathbf{u}_l'\mathbf{u}_{l'}=0$ and $\mathbf{u}_l'\mathbf{u}_l=1$, hence $\mathbf{u}_1,\ldots,\mathbf{u}_r$ can be seen as the columns of a columnwise orthonormal matrix **U**. To maximize (14) we can use the following iterative algorithm:

```
Step 1. Initialize $\mathbf{U}^i$ ($i$=0) (e.g., as random columnwise orthonormal matrix)
Step 2. Compute h($\mathbf{U}^i$)
Step 3
    a. Compute $\mathbf{G}_l = \mathbf{S}\mathbf{u}_l^i\mathbf{u}_l^{i\prime}\mathbf{T}+\mathbf{T}\mathbf{u}_l^i\mathbf{u}_l^{i\prime}\mathbf{S}$, $l$=1,...,$r$.
    b. Compute a matrix F with columns $\mathbf{G}_l\mathbf{u}_l^i-\mu_l\mathbf{u}_l^i$, where $\mu_l$ is the smallest
       eigenvalue of $\mathbf{G}_l$
    c. Compute the SVD $\mathbf{F}=\mathbf{PDQ}'$, and compute $\mathbf{U}^{i+1}=\mathbf{PQ}'$
Step 4. Compute h($\mathbf{U}^{i+1}$); if h($\mathbf{U}^{i+1}$)–h($\mathbf{U}^i$) > $\varepsilon$h($\mathbf{U}^i$) for some prespecified small
    value $\varepsilon$ (e.g., $\varepsilon=10^{-6}$), then go to Step 3; else consider the algorithm
    converged.
```

The above algorithm increases h(**U**) monotonically, and because h($\mathbf{U}^i$) is bounded, the algorithm converges to a stable function value. A proof for the monotonicity of the algorithm is not given here, but is based on the majorisation algorithm offered by Kiers (1990). It should be noted that the above algorithm can be used for maximizing (12) and (13). If it is used for maximizing (13), it can be proven that $\tilde{\mathbf{t}}_l$ remains in the columns space of **X′X**.

Using the above algorithm, we solve indirectly for $\mathbf{w}_l$, $l=1,\ldots,r$, subject to the constraint $\mathbf{w}_l'\mathbf{X}'\mathbf{X}\mathbf{w}_{l'}=0$ if $l'\neq l$. Thus, we can obtain the mutually orthogonal components $\mathbf{X}\mathbf{w}_1,\ldots,\mathbf{X}\mathbf{w}_r$ and regress the criterion scores in **Y** onto these components; note that, due the mutual orthogonality of these predictors, this can be done for each component separately, yielding the matrix of regression weights $\mathbf{P}_\mathbf{Y}$. Finally, the regression weights matrix **B** is computed as $\mathbf{B} = \mathbf{W}\mathbf{P}_\mathbf{Y}$, thus yielding the prediction rule for estimating the scores on **Y** from those on **X**.

The above algorithm can be used for $r\geq1$. For r=1, this algorithm is not identical to the one given in Section 4.1. The latter is easier to program.

5. Three-way Extensions of PcovR and Power regression

In the last decade, there has been a growing interest in more complex data types, like three-way data sets and multi-block data sets. If a three-way array of data is to be used to predict outcomes of a (set of) criterion variable(s), one could consider the three-way data set as a big two-way data set with scores of observation units on combinations of, for instance, variables and conditions. Then, obviously, the wish to summarize all these predictor variables becomes rather strong. Smilde (1997), see also Smilde and Kiers (1999) proposed a variant of PCovR to summarize the three-way array of predictor variables by a three-way model, combined with the prediction of criterion variables. With Power Regression, in principle a similar approach could be followed, but algorithms for maximizing the ensuing criteria still remain to be developed.

6. Conclusion

In the present paper several regression methods have been presented that do not only aim at explaining a large amount of variance of the criterion variable(s), but also of the predictor variables. The rationale for this approach is that by aiming at explaining a large amount of variance of the predictor variables (in addition to that of the criterion variable(s)) will lead to regression weights that are not only optimal for the training data set, but will also work well for other data sets. One of the first methods that aimed (implicitly) at this goal was PLS, which indeed tends to perform well in practice. The alternatives presented here are aimed *explicitly* at this goal of explaining variance of the criterion and the predictor variables. It has been seen in test analyses that these methods can give solutions that explain more variance in both the criterion and the predictor variables than PLS does. Also, it has been found in a simulation that such methods can give weights that are closer to the ‘true’ regression weights than PLS does. However, experience with these methods is still limited, and little can be said, as yet, on the performance of the methods in actual practice. Clearly, further research is needed to test the new methods and compare them to each other and to PLS.

7. Appendix: proof for the monotonicity of the algorithm in 4.1

It is to be proven that for h defined in (9), we have $h(\mathbf{v}) \geq h(\mathbf{u})$ when $\mathbf{v}$ is chosen as the first eigenvector of $\mathbf{Suu'T+Tuu'S}$, and $\mathbf{u'u} = \mathbf{v'v} =1$. From the fact that $\mathbf{v}$ is the first eigenvector of $\mathbf{Suu'T+Tuu'S}$, we have $2(\mathbf{v'Su})(\mathbf{v'Tu}) = \mathbf{v'Suu'Tv} + \mathbf{v'Tuu'Sv} \geq \mathbf{u'Suu'Tu} + \mathbf{u'Tuu'Su} = 2\mathbf{u'Suu'Tu}$. Furthermore, we have $(\mathbf{v'Sv})^{1/2}(\mathbf{u'Su})^{1/2} \geq \mathbf{v'Su}$ from the Cauchy-Schwarz inequality, and, analogously, $(\mathbf{v'Tv})^{1/2}(\mathbf{u'Tu})^{1/2} \geq \mathbf{v'Tu}$. Combining these inequalities, we have $2(\mathbf{v'Sv})^{1/2}(\mathbf{v'Tv})^{1/2}(\mathbf{u'Tu})^{1/2}(\mathbf{u'Su})^{1/2} \geq 2(\mathbf{v'Su})(\mathbf{v'Tu}) \geq 2\mathbf{u'Suu'Tu'}$, hence $(\mathbf{v'Svv'Tv})^{1/2}(\mathbf{u'Suu'Tu})^{1/2} \geq \mathbf{u'Suu'Tu'}$. From this it follows that $(\mathbf{v'Svv'Tv})^{1/2} \geq (\mathbf{u'Suu'Tu})^{1/2}$, hence $h(\mathbf{v}) = \mathbf{v'Svv'Tv} \geq \mathbf{u'Suu'Tu} = h(\mathbf{u})$.

References

Breiman, L., Friedman, J.H. (1997) Predicting multivariate responses in multiple linear regression (with discussion), *Journal of the Royal Statistical Society, Series B*, 59, 3-54.

Coxe, K.L. (1986) Principal components regression analysis, in: *Encyclopedia of Statistical Sciences, Vol. 7*, N.L. Johnson & Kotz, S. (Eds.), Wiley, New York, 181-186.

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

De Jong, S., Kiers, H.A.L. (1992) Principal Covariates Regression: Part I. Theory, *Chemometrics and Intelligent Laboratory Systems,* 14, 155-164.

Esposito Vinzi, V., Lauro, C., Morineau, A., Tenenhaus, M. (2001) *PLS and related methods. Proceedings of the PLS'01 International Symposium*, CISIA-CERESTA, Montreuil, France.

Frank, I.E. (1987) Intermediate least squares regression method, *Chemometrics and Intelligent Laboratory Systems,* 1, 233-242.

Hoerl, A.E., Kennard, R.W. (1970) Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, 12, 55-67.

Kiers, H.A.L. (1990) Maximizaiton as a tool for optimizing a class of matrix functions, *Psychometrika,* 55, 417-428.

Manne, R. (1987) Analysis of partial-least-squares algorithms for multivariate calibration, *Chemometrics and Intelligent Laboratory Systems,* 2, 283-290.

Martens, H., Naes, T. (1989) *Multivariate Calibration*, Wiley, New York.

Smilde, A.K. (1997) Comments on multilinear PLS, *Journal of Chemometrics, 11,* 367-377.

Smilde, A.K., Kiers, H.A.L. (1999) Multiway covariates regression models. *Journal of Chemometrics, 13,* 31-48.

Stone, M., Brooks, R.J. (1990) Continuum regression: cross-validated sequentially constructred prediction embracing ordinary least squares, partial least squares and principal components regression (with discussion), *Journal of the Royal Statistical Society, Series B*, 52, 237-269.

Wold, H. (1966) Estimation of principal components and related models by iterative least squares, in: *Multivariate Analysis*, P.R. Krishnaiah (Ed.), Academic press, New York, 391-420.

Wold, S., Ruhe, A., Wold, H., Dunn III, W.J. (1984) The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses, *SIAM Journal of Scientific and Statistical Computing,* 5, 735-744.