Reduced PCR/PLSR models by subspace projections

Rolf Ergon
Telemark University College
P.O.Box 203, N-3901 Porsgrunn, Norway
e-mail: rolf.ergon@hit.no

Published in Chemometrics and Intelligent Laboratory Systems 81 (2006) 68-73

Abstract

Latent variables models used in principal component regression (PCR) or partial least squares regression (PLSR) often use a high number of components, and this makes interpretation of score and loading plots difficult. These plots are essential parts of multivariate modeling, and there is therefore a need for a reduction of the number of components without loss of prediction power. In this work it is shown that such reductions of PCR models with a common number of components for all responses, as well as of PLSR (PLS1 and PLS2) models, may be obtained by projection of the $\mathbf{X}$ modeling objects onto a subspace containing the estimators $\hat{\mathbf{b}}_i$ for the different responses $\mathbf{y}_i$. The theoretical results are substantiated in three real world data set examples, also showing that the presented model reduction method may work quite well also for PCR models with different numbers of components for different responses, as well as for a set of individual PLSR (PLS1) models. Examples of interpretational advantages of reduced models in process monitoring applications are included.

Keywords: PCR; PLSR; Model reduction; Subspace projection

1 Introduction

Solutions of ill-posed regression problems using principal component regression (PCR) or partial least squares regression (PLSR) based on latent variables (LV) models, offer the added advantage of score-loading visualizations. However, for good predictions such models often require so many components that interpretation of the plots is difficult, and this calls for model reduction methods without loss of prediction capability. This is the main results of the orthogonal signal correction (OSC) methods [1], and as pointed out by Trygg [2] the number of final components need not be higher than the number of independent responses. This fact also follows from an earlier paper by Kvalheim and Karstang [3]. With a single response, one component will thus be enough, but that will obviously not facilitate the score and loading plots that are essential parts of multivariate modeling. This was the motivation behind the 2PLS algorithm [4], where a single response LV model (PLS1) with many components is reduced to a model with two components only. That is obtained by projection of the objects in the A-dimensional loading weights space onto a two-dimensional subspace containing the vector $\hat{\mathbf{b}}$ of prediction coefficients (the estimator).

In the present paper the 2PLS theory is extended to show that any PCR or PLSR (PLS1 and PLS2) model with an A-dimensional loading space can be reduced by projection onto a subspace with any dimension $m \leq \tilde{A} < A$, where m is the number of independent responses, and that this can be done while retaining exactly the same predictions. Since the only significant difference between PCR and PLSR (especially PLS1) often is the number of original components, and since this difference disappears after the model reduction, the two methods will often give practically the same final results. The model reduction results may utilized in e.g. operator support systems for process understanding and monitoring [5,6].

In the multiresponse case with $m > 1$ the corresponding reduction of a set of individual PLSR (PLS1) models will not give exactly the same predictions, and the reason for this is that the individual coefficient vectors in $\hat{\mathbf{B}}_{\text{PLS1}}$ are located in different subspaces. The same problem will occur for a set of individual PCR estimators based on different numbers of components. The predictions in these cases may be approximately the same, however, just as a multiresponse PLS2 models may give approximately the same predictions as a set of individual single response PLS1 models.

The theory behind these model reductions is given in Section 2, while comparisons between the different cases using real world data are presented in Section 3. Section 3 also includes illustrations of the interpretational advantages obtained. Conclusions are finally given in Section 4, and a detailed proof in Appendix A.

2 Theory

LV models

Assume an ill-posed linear regression problem with modeling data $\mathbf{X} \in \mathbb{R}^{N\times p}$ and $\mathbf{Y} \in \mathbb{R}^{N\times m}$, and solutions using PCR or PLSR. For simplicity also assume that $\mathbf{Y}$ is non-collinear, or alternatively replaced by an appropriate number of principal components of $\mathbf{Y}$. In PCR we use the LV model

$$\begin{aligned}\mathbf{Y} &= \mathbf{T}\mathbf{Q}^T + \mathbf{F} & (1)\\ \mathbf{X} &= \mathbf{T}\mathbf{P}^T + \mathbf{E}, & (2)\end{aligned}$$

where both $\mathbf{T} \in \mathbb{R}^{N\times A}$ and $\mathbf{P} \in \mathbb{R}^{p\times A}$ are orthogonal, and $\mathbf{P}$ also orthonormal. Least squares (LS) estimation gives $\hat{\mathbf{T}} = \mathbf{X}\mathbf{P}\left(\mathbf{P}^T\mathbf{P}\right)^{-1} = \mathbf{X}\mathbf{P}$ and $\hat{\mathbf{Q}}^T = \left(\hat{\mathbf{T}}^T\hat{\mathbf{T}}\right)^{-1}\hat{\mathbf{T}}^T\mathbf{Y}$, and the estimator (matrix of prediction coefficients) in $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}} = \hat{\mathbf{T}}\hat{\mathbf{Q}}^T$ thus becomes

$$\hat{\mathbf{B}}_{\text{PCR}} = \mathbf{P}\left(\mathbf{P}^T\mathbf{X}^T\mathbf{X}\mathbf{P}\right)^{-1}\mathbf{P}^T\mathbf{X}^T\mathbf{Y}. \qquad (3)$$

The loading matrix $\mathbf{P}$ is here found by principal component analysis (PCA) or singular value decomposition (SVD) of $\mathbf{X}$, with the number of components determined through validation. This ensures that the individual estimators (columns of $\hat{\mathbf{B}}_{\text{PCR}}$) are all located in the column space of $\mathbf{P}$. This is the case also if the individual estimators are found one by one, provided that the *same number of components* is used in all individual PCR models.

In PLSR we may use the orthogonalized LV model

$$\begin{aligned}\mathbf{Y} &= \mathbf{T}_{\text{W}}\mathbf{Q}_{\text{W}}^T + \mathbf{F} & (4)\\ \mathbf{X} &= \mathbf{T}_{\text{W}}\mathbf{P}_{\text{W}}^T\mathbf{W}\mathbf{W}^T + \mathbf{E}, & (5)\end{aligned}$$

possibly with $\mathbf{X}$ expressed as $\mathbf{X} = \mathbf{T}_{\text{W}}\mathbf{P}_{\text{W}}^T + \mathbf{E}_{\text{W}}$, where both $\mathbf{T}_{\text{W}} \in \mathbb{R}^{N\times A}$ and $\mathbf{W} \in \mathbb{R}^{p\times A}$ are orthogonal, and $\mathbf{W}$ also orthonormal. Alternatively we may use the non-orthogonalized LV model

$$\begin{aligned}\mathbf{Y} &= \mathbf{T}_{\text{M}}\mathbf{Q}_{\text{M}}^T + \mathbf{F} & (6)\\ \mathbf{X} &= \mathbf{T}_{\text{M}}\mathbf{W}^T + \mathbf{E}, & (7)\end{aligned}$$

where $\mathbf{T}_{\text{M}}$ is non-orthogonal. From all of these PLSR models follows the LS estimator

$$\hat{\mathbf{B}}_{\text{PLSR}} = \mathbf{W}\left(\mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W}\right)^{-1}\mathbf{W}^T\mathbf{X}^T\mathbf{Y}. \qquad (8)$$

The loading weights matrix $\mathbf{W}$ is here found by the step-wise NIPALS algorithm, with some extensions for the multiresponse case [7]. PLSR differs from PCR in that the columns of $\hat{\mathbf{B}}_{\text{PLSR}}$ in the multiresponse case (PLS2) *are not the same* as the individual estimators for the different single response cases. Individual PLSR for different responses y_i (PLS1) will result in different loading weights matrices $\mathbf{W}_i$, and the different estimators $\hat{\mathbf{b}}_i$ will therefore be located in different subspaces. A common PLSR for all responses (PLS2) will ensure individual estimators in the same subspace, often at the cost of degraded predictions.

Finally note that in the case of collinear responses, the matrix $\mathbf{Y}$ may be replaced by an appropriate number of principal components of $\mathbf{Y}$.

Model reduction

In PCR and PLSR all objects in $\mathbf{X}$ are projected onto the column space of $\mathbf{P}$ or $\mathbf{W}$. They may also be further projected onto subspaces of these column spaces, e.g. the column space of $\left[\begin{array}{cc}\hat{\mathbf{B}}_{\mathrm{PCR}} & \mathbf{PL}\end{array}\right] \in \mathbb{R}^{p\times\tilde{A}}$ or $\left[\begin{array}{cc}\hat{\mathbf{B}}_{\mathrm{PLSR}} & \mathbf{WL}\end{array}\right] \in \mathbb{R}^{p\times\tilde{A}}$, where $m \leq \tilde{A} < A$, and where $\mathbf{PL}$ or $\mathbf{WL}$ are linear combinations of $\mathbf{P}$ or $\mathbf{W}$ that are not collinear with $\hat{\mathbf{B}}_{\mathrm{PCR}}$ or $\hat{\mathbf{B}}_{\mathrm{PLSR}}$. After such a projection we may define an orthonormal matrix $\tilde{\mathbf{P}}$ spanning such a subspace, and an LV model

$$\mathbf{Y} = \tilde{\mathbf{T}}\tilde{\mathbf{Q}}^T + \tilde{\mathbf{F}} \tag{9}$$

$$\mathbf{X} = \tilde{\mathbf{T}}\tilde{\mathbf{P}}^T + \tilde{\mathbf{E}}, \tag{10}$$

where $\tilde{\mathbf{T}} = \mathbf{X}\tilde{\mathbf{P}}$ and $\tilde{\mathbf{Q}}^T = \left(\tilde{\mathbf{T}}^T\tilde{\mathbf{T}}\right)^{-1}\tilde{\mathbf{T}}^T\mathbf{Y}$. In the same way as for the estimators (3) and (8) this will result in the LS estimator

$$\tilde{\mathbf{B}} = \tilde{\mathbf{P}}\left(\tilde{\mathbf{P}}^T\mathbf{X}^T\mathbf{X}\tilde{\mathbf{P}}\right)^{-1}\tilde{\mathbf{P}}^T\mathbf{X}^T\mathbf{Y}, \tag{11}$$

and assuming a PCR model with the same number of components for all responses as a starting point, this estimator will be exactly the same as the PCR estimator (3). This is true also for a single response PLSR model (PLS1), and for a common multiresponse PLSR model (PLS2). We summarize in the following theorem:

Theorem 1 *A PCR, PLS1 or PLS2 latent variables model with m independent response variables and A components, can by projection of all modeling objects in $\mathbf{X}$ onto a subspace containing the columns of the estimator $\hat{\mathbf{B}}_{PCR}$ or $\hat{\mathbf{B}}_{PLSR}$ respectively be reduced to an LV model with $m \leq \tilde{A} < A$ components, while still retaining exactly the same estimator.*

Proof. See Appendix A. ■

Note that Theorem 1 is not valid for a set of individual PCR models with different numbers of components, or for a set of individual PLSR (PLS1) models. The reason for this is that the individual estimators $\hat{\mathbf{b}}_i$ in these cases generally are located in different subspaces. Also in these cases, however, we may obtain results that are approximately in accordance with the theorem (see examples in Section 3 below).

3 Examples

The main point with the examples below is to indicate the possibility to obtain approximate results according to Theorem 1 also in the multiple response PCR and PLSR cases where the reduced model estimators are not exactly equal to the original set of individual estimators. Analogously to PLSR we will here use the notation PCR1 for individual single response PCR results, and PCR2 for multiresponse PCR with a common number of components. Examples of interpretational advantages of model reduction are given in the references [4,5,6], but for the readers convenience some illustrative results from [5] are also included here.

3.1 Gasoil example

The data in this example are provided by the Wentzell group at Dalhousie University (http://myweb.dal.ca/pdwentze/downloads.html, Data Set #3), under the name "gasoil". The data set consists of 115 samples for which the UV spectra over 572 channels have been obtained. The data includes four response variables. After centering and standardization of the data, and using the first 70 samples for modeling and the next 44 samples for validation (the last is an outlier), the results in Table 1 were obtained.

Table 1: RMSE values for the four Gasoil responses using different LV models. A common "best possible" value of A was chosen for the original PCR2 and PLS2 estimators, while the PCR1 and PLS1 estimators were individually optimized. The best result for each response is underlined.

	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$\hat{y}_4$
PCR2 model with $A = 15$, and reduced PCR2 model with $\tilde{A} = 4$	0.1087	0.1477	0.1851	0.2764
Individual PCR1 models with $A = 23$, 23, 23 and 15	0.1026	0.1414	0.1824	0.2764
Reduced PCR2 model with $\tilde{A} = 4$, based on individual PCR1 models with $A = 23$, 23, 23 and 15	0.1026	0.1414	0.1824	0.2759
Individual PLS1 models with $A = 6$, 5, 4 and 4	0.1129	0.1434	0.1749	0.2774
Reduced PLS2 model with $\tilde{A} = 4$, based on individual PLS1 models with $A = 6$, 5, 4 and 4	0.1109	0.1384	0.1989	0.2838
PLS2 model with $A = 5$, and reduced PLS2 model with $\tilde{A} = 4$	0.1215	0.1471	0.1615	0.2883

From Table 1 it is clear that none of the methods is decisively better than the others. In accordance with the theory, reduction of the PCR2 and PLS2 models give exactly the same results as the original ones, while reduction of the individual PCR1 and PLS1 models give results slightly different from the original ones (same PCR2 result for response with highest number of PCR1 components, since the subspaces then coincide).

3.2 Cleaner example

The following example uses multivariate regression data from a mineral processing plant [8] (the 'cleaner' data, originally published in [9]). The problem considered here is to predict two given responses y_4 and y_7 from twelve known process variables.

Prediction results

After centering and standardization of the data, and using the first 120 samples for modeling and samples 181-240 for validation, the results in Table 2 were obtained.

Table 2: RMSE values for the two Cleaner responses in different cases. A common "best possible" value of A was chosen for the PCR2 and PLS2 estimators, while the PCR1 and PLS1 estimators were individually optimized. The best result for each response is underlined.

	$\hat{y}_4$	$\hat{y}_7$
PCR2 model with $A = 10$, and reduced PCR2 model with $\tilde{A} = 4$	0.1880	0.2783
Individual PCR1 models with $A = 10$ and 10	0.1880	0.2783
Reduced PCR2 model with $\tilde{A} = 4$, based on individual PCR1 models with $A = 10$ and 10	0.1880	0.2783
Individual PLS1 models with $A = 6$ and 3	0.1901	0.2804
Reduced PLS2 model with $\tilde{A} = 4$, based on individual PLS1 models with $A = 6$ and 3	0.1930	0.2799
PLS2 model with $A = 6$, and reduced PLS2 model with $\tilde{A} = 2$	0.2463	0.2938

The conclusions drawn from Table 2 are essentially the same as in the Gasoil example above.

Interpretational illustrations

In order to illustrate the interpretational advantages the PLS1 model (6,7) for response y_4 using $A = 6$ components is reduced to a 2PLS model [4] by projection onto the plane defined by the estimator $\hat{\mathbf{b}}$ and the first loading weight vector $\mathbf{w}_1$, resulting in the LV model (9,10). New $\mathbf{X}$ data were then introduced as $\mathbf{X}^{\text{test}} = \mathbf{I}_{12}$, i.e.

$$\mathbf{X}^{\text{test}} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{12}$$

The 2PLS scores and loadings for $\tilde{\mathbf{T}}^{\text{test}} = \mathbf{X}^{\text{test}}\tilde{\mathbf{P}}$ and $\tilde{\mathbf{P}}$ were then plotted (Fig. 1). The loadings were here normalized to fall on the unit circle ($*$-marked). The scores corresponding to $\mathbf{X}^{\text{test}}$ are also plotted in Fig. 1 (×-marked). As can be seen, the direction for e.g. test score number 2 corresponds exactly to the direction for row number 2 in the $\tilde{\mathbf{P}}$ matrix, i.e. there is an exact score-loading correspondence (see [10] for a general discussion). In a process monitoring application a large deviation from the origin thus not only signals a special plant operation situation, but also indicates which regressor variable or variables that cause the deviation.

Fig. 1 also shows lines for constant $\hat{y}$ and an axis for $\hat{\mathbf{b}}$ and thus $\hat{y}$ perpendicular to those lines. The plot can be used to gain process understanding. As can be seen the predicted response $\hat{y}$ is strongly correlated with variable 3 and also with variables 1, 2, 8 and 9, while the other variables according to the loading directions have little to do with $\hat{y}$ (and variable 9 has a very limited influence). For process monitoring purposes, variable contribution vectors and a confidence ellipse may also be incorporated in the score loading biplot [5,6].

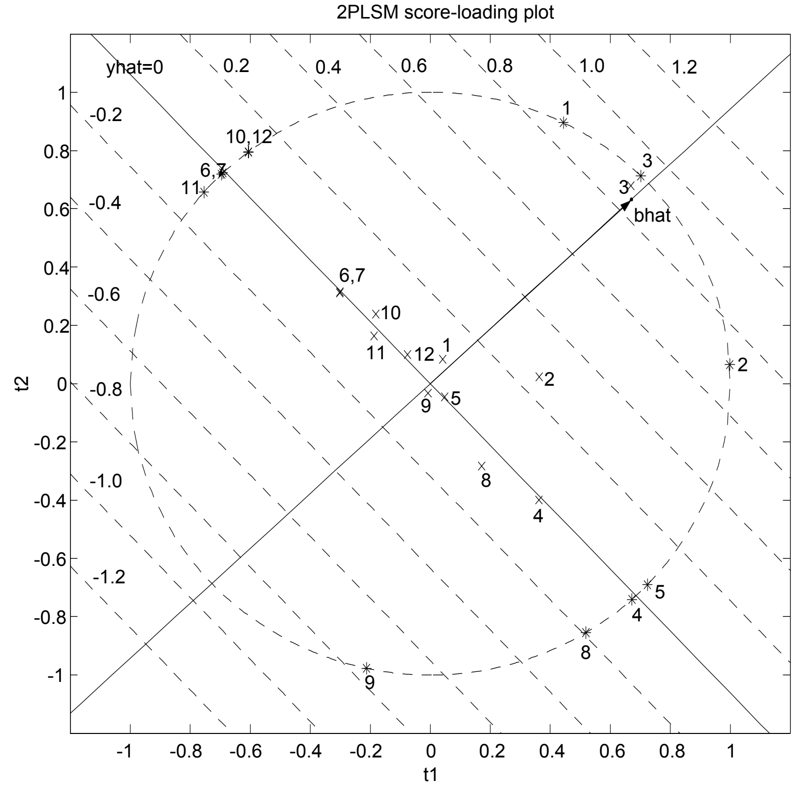

Figure 1. Score-loading plot for $\mathbf{X}^{\text{test}}$ according to Eq. (12), based on a reduced model with $\tilde{A} = 2$ components (2PLS). The scores are ×-marked, while the normalized loadings are $*$ -marked. Dashed lines for constant predictions $\hat{y}$ are included in the score plot, while the estimator $\hat{\mathbf{b}}$ is included in the loading plot. Note

the exact score-loading correspondence, and that $\hat{\mathbf{b}}$ is perpendicular to the loadings 4, 5 and 6 (compare with Fig. 2).

Using original score plots and projections of $\hat{\mathbf{b}}$ onto the corresponding loading plots, we obtain the results in Fig. 2. The prediction $\hat{y}$ cannot in this case be captured in any of these plots, and the influence from the different variables on the scores are thus obscured. It is for example not possible to see that $\hat{\mathbf{b}}$ is perpendicular to a plane through the test scores 4, 5 and 6, as can be seen from Fig. 1, and thus that $\hat{y}$ is not influenced by the corresponding variables. This can of course be seen directly from the $\hat{\mathbf{b}}$ coefficients, but that hardly helps an operator in the interpretation of the plots.

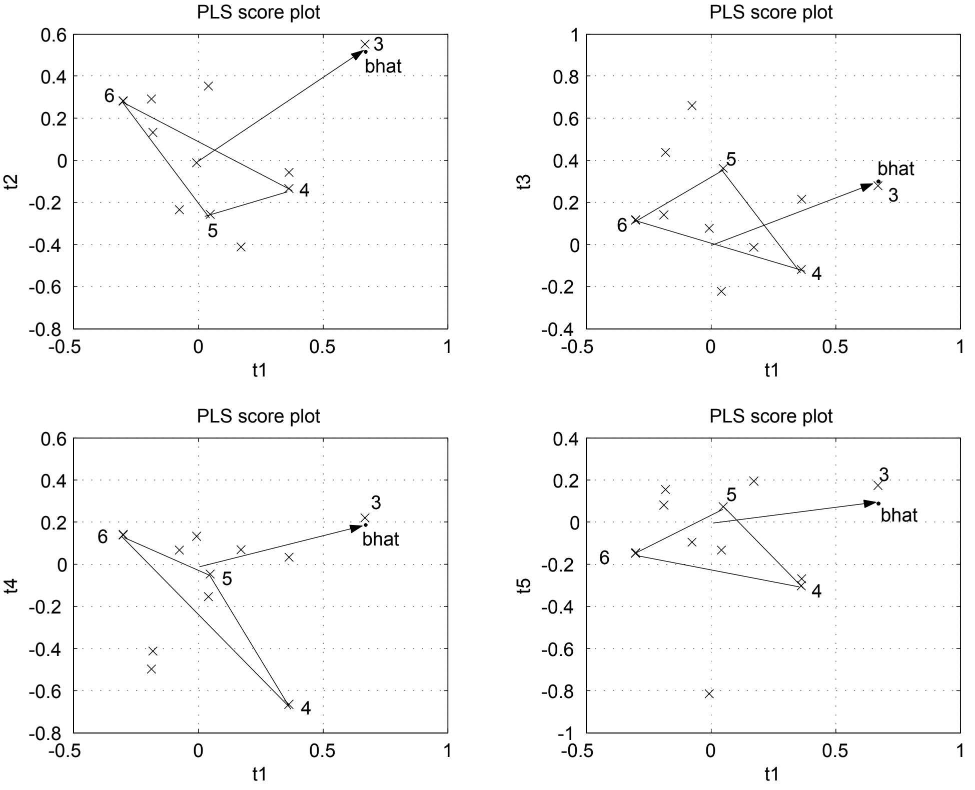

Figure 2. Score plot for $\mathbf{X}^{\text{test}}$ according to Eq. (12) with ×-marked scores, based on an original PLS1 model with $A = 6$ components. Projections of the estimator $\hat{\mathbf{b}}$ onto the corresponding loading plots are superimposed. Note that the direction of $\hat{y}$ cannot be seen in these plots. Neither can it be seen that $\hat{\mathbf{b}}$ is perpendicular to the loadings 4, 5 and 6 (compare with Fig. 1).

3.3 Corn example

This data originating from the Cargill company is found on the Eigenvector home site (http://software.eigenvector.com/Data/Corn/index.html). It consists of 80 samples of corn measured on a NIR spectrometer labeled m5. The wavelength range is 1100-2498 nm at 2 nm intervals (700 channels). The moisture (y_1), oil (y_2), protein (y_3) and starch (y_4) values for each of the samples are also included. After centering and standardization of the data, and using the first 40 samples for modeling and samples 41-80 for validation, the results in Table 3 were obtained.

Table 3: RMSE values for the four Corn responses in different cases. A common "best possible" value of A was chosen for the PCR2 and PLS2 estimators, while the PCR1 and PLS1 estimators were individually optimized. The best result for each response is underlined.

	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$\hat{y}_4$
PCR2 model with $A = 26$, and reduced PCR2 model with $\tilde{A} = 4$	0.3609	0.9732	0.4420	0.4885
Individual PCR1 models with $A = 27$, 25, 26 and 25	0.3485	0.9536	0.4420	0.4443
Reduced PCR2 model with $\tilde{A} = 4$, based on individual PCR1 models with $A = A = 27$, 25, 26 and 25	0.3485	0.9518	0.4420	0.4467
Individual PLS1 models with $A = 12$, 17, 8 and 9	0.3506	0.6912	0.4466	0.5010
Reduced PLS2 model with $\tilde{A} = 4$, based on individual PLS1 models with $A = 1$, 17, 8 and 9	0.3495	0.6902	0.4349	0.5054
PLS2 model with $A = 13$, and reduced PLS2 model with $\tilde{A} = 4$	0.3551	1.0128	0.4649	0.5198

In Table 3 we see that for variables y_1 (moisture) and y_3 (protein) all methods give very similar results. For variable y_2 (oil) only the individual PLS1 models and the PLS2 model based on these give in any way useful estimates, and the reduced model is just as good as the individual ones. For variable y_4 (starch), the PCR estimators are clearly the best. Note, however, that these results are based on a relatively limited number of samples.

4 Conclusions

The main idea behind the model reduction results presented above is that models for prediction and interpretation should not necessarily be the same, and the practical result of model reduction is increased interpretability. An example of that is the 2PLS algorithm, where a single response PLSR model is reduced by projection on a plane containing $\hat{\mathbf{b}}$ [4].

It has been shown that a PCR latent variables model with a common number of components for all responses, as well as a PLSR model for one or several responses (PLS1 or PLS2), can without any change in the predictions be reduced by projection of the $\mathbf{X}$ modeling objects onto a subspace containing the estimators $\hat{\mathbf{b}}_i$ for the different responses $\mathbf{y}_i$. The result of this is easier interpretation of the score and loading plots, that are essential parts of multivariate modeling.

The theoretical results are substantiated by use of three real world data set examples, also showing that the presented model reduction method may work quite well also for a set of individual PCR latent variables models with different numbers of components for different responses, as well as for a set of individual PLSR (PLS1) models. Some interpretational advantages are also exemplified, and more details of this are found in the references.

A Proof of Theorem 1

The proof is given for the PCR estimator (3) based on a latent variables model with loading matrix $\mathbf{P}$. For a PLSR estimator we must generally replace $\mathbf{P}$ with $\mathbf{W}$, and for a PLS1 model we must assume $m = 1$.

We first prove the theorem for the special case of $\tilde{A} = m$. Using the simplified notation $\hat{\mathbf{B}}_{\text{PCR}} = \hat{\mathbf{B}}$, and introducing an invertible matrix $\mathbf{M} \in \mathbb{R}^{m \times m}$ and an orthonormal loading matrix $\tilde{\mathbf{P}} = \hat{\mathbf{B}}\mathbf{M}$ with the same column space as $\hat{\mathbf{B}}$, we obtain by use of Eq. (3) and with $(\cdot)^{-1} = (\mathbf{P}^T\mathbf{X}^T\mathbf{X}\mathbf{P})^{-1}$

$$\begin{aligned}
\tilde{\mathbf{B}} &= \tilde{\mathbf{P}}(\tilde{\mathbf{P}}^T\mathbf{X}^T\mathbf{X}\tilde{\mathbf{P}})^{-1}\tilde{\mathbf{P}}^T\mathbf{X}^T\mathbf{Y} = \hat{\mathbf{B}}\mathbf{M}(\mathbf{M}^T\hat{\mathbf{B}}^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{B}}\mathbf{M})^{-1}\mathbf{M}^T\hat{\mathbf{B}}^T\mathbf{X}^T\mathbf{Y} \\
&= \hat{\mathbf{B}}(\hat{\mathbf{B}}^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^T\mathbf{X}^T\mathbf{Y} \\
&= \mathbf{P}(\cdot)^{-1}\mathbf{P}^T\mathbf{X}^T\mathbf{Y}\left(\mathbf{Y}^T\mathbf{X}\mathbf{P}\,(\cdot)^{-1}\,\mathbf{P}^T\mathbf{X}^T\mathbf{X}\mathbf{P}\,(\cdot)^{-1}\,\mathbf{P}^T\mathbf{X}^T\mathbf{Y}\right)^{-1}\mathbf{Y}^T\mathbf{X}\mathbf{P}\,(\cdot)^{-1}\,\mathbf{P}^T\mathbf{X}^T\mathbf{Y} \\
&= \mathbf{P}(\cdot)^{-1}\mathbf{P}^T\mathbf{X}^T\mathbf{Y} = \mathbf{P}(\mathbf{P}^T\mathbf{X}^T\mathbf{X}\mathbf{P})^{-1}\mathbf{P}^T\mathbf{X}^T\mathbf{Y} = \hat{\mathbf{B}}. \qquad (13)
\end{aligned}$$

Note that $\mathbf{M} = (\hat{\mathbf{B}}^T\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^T\tilde{\mathbf{P}}$, where the actual choice of $\tilde{\mathbf{P}}$ and thus $\mathbf{M}$ is application dependent. In any case we see that the columns of $\hat{\mathbf{B}}$ and thus of $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$ must be independent, as stated in Theorem 1.

For the general case allowing for $\tilde{A} > m$ we introduce a matrix $\mathbf{L} \in \mathbb{R}^{A\times(\tilde{A}-m)}$ and use $\tilde{\mathbf{P}} = \left[\begin{array}{cc}\hat{\mathbf{B}}\mathbf{M} & \mathbf{P}\mathbf{L}\end{array}\right]$, where $\mathbf{L}$ is chosen such that the column space of $\left[\begin{array}{cc}\hat{\mathbf{B}}\mathbf{M} & \mathbf{P}\mathbf{L}\end{array}\right]$ is of dimension $\tilde{A}$. The actual choices of $\mathbf{M}$ and $\mathbf{L}$ will here depend on the application at hand. Using $\hat{\mathbf{B}} = \mathbf{B}$ and introducing $\mathbf{R} = \left(\mathbf{M}^T\mathbf{B}^T\mathbf{X}^T\mathbf{X}\mathbf{B}\mathbf{M}\right)^{-1}$ for simplified notation, the resulting estimator then becomes

$$\begin{aligned}
\tilde{\mathbf{B}} &= \left[\begin{array}{cc}\mathbf{B}\mathbf{M} & \mathbf{P}\mathbf{L}\end{array}\right]\left(\left[\begin{array}{c}\mathbf{M}^T\mathbf{B}^T \\ \mathbf{L}^T\mathbf{P}^T\end{array}\right]\mathbf{X}^T\mathbf{X}\left[\begin{array}{cc}\mathbf{B}\mathbf{M} & \mathbf{P}\mathbf{L}\end{array}\right]\right)^{-1}\left[\begin{array}{c}\mathbf{M}^T\mathbf{B}^T \\ \mathbf{L}^T\mathbf{P}^T\end{array}\right]\mathbf{X}^T\mathbf{Y} \\
&= \left(\begin{array}{c}
\mathbf{B}\mathbf{M}\mathbf{R}\mathbf{M}^T\mathbf{B}^T + \mathbf{B}\mathbf{M}\mathbf{R}\mathbf{M}^T\mathbf{B}^T\mathbf{X}^T\mathbf{X}\mathbf{P}\mathbf{L}\,[\cdot]^{-1}\,\mathbf{L}^T\mathbf{P}^T\mathbf{X}^T\mathbf{X}\mathbf{B}\mathbf{M}\mathbf{R}^T\mathbf{M}^T\mathbf{B}^T \\
-\mathbf{P}\mathbf{L}\,[\cdot]^{-1}\,\mathbf{L}^T\mathbf{P}^T\mathbf{X}^T\mathbf{X}\mathbf{B}\mathbf{M}\mathbf{R}^T\mathbf{M}^T\mathbf{B}^T \\
-\mathbf{B}\mathbf{M}\mathbf{R}\mathbf{M}^T\mathbf{B}^T\mathbf{X}^T\mathbf{X}\mathbf{P}\mathbf{L}\,[\cdot]^{-1}\,\mathbf{L}^T\mathbf{P}^T + \mathbf{P}\mathbf{L}\,[\cdot]^{-1}\,\mathbf{L}^T\mathbf{P}^T
\end{array}\right)\mathbf{X}^T\mathbf{Y} \qquad (14) \\
&= \mathbf{B}\mathbf{M}\mathbf{R}\mathbf{M}^T\mathbf{B}^T\mathbf{X}^T\mathbf{Y} + \left(\mathbf{B}\mathbf{M}\mathbf{R}\mathbf{M}^T\mathbf{B}^T\mathbf{X}^T\mathbf{X}\mathbf{P}\mathbf{L} - \mathbf{P}\mathbf{L}\right)[\cdot]^{-1}\left(\mathbf{B}\mathbf{M}\mathbf{R}\mathbf{M}^T\mathbf{B}^T\mathbf{X}^T\mathbf{X}\mathbf{P}\mathbf{L} - \mathbf{P}\mathbf{L}\right)^T\mathbf{X}^T\mathbf{Y},
\end{aligned}$$

where a standard formula for matrix inversion [11] comes to use, and where $[\cdot]^{-1} = \left[\mathbf{L}^T\mathbf{P}^T\left(\mathbf{X}^T\mathbf{X} - \mathbf{X}^T\mathbf{X}\mathbf{B}\mathbf{M}\mathbf{R}\mathbf{M}^T\mathbf{B}^T\mathbf{X}^T\mathbf{X}\right)\mathbf{P}\mathbf{L}\right]^{-1}$. But using that $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{P}\left(\mathbf{P}^T\mathbf{X}^T\mathbf{X}\mathbf{P}\right)^{-1}\mathbf{P}^T\mathbf{X}^T\mathbf{Y} = \mathbf{X}\mathbf{B}\mathbf{M}\left(\mathbf{M}^T\mathbf{B}^T\mathbf{X}^T\mathbf{X}\mathbf{B}\mathbf{M}\right)^{-1}\mathbf{M}^T\mathbf{B}^T\mathbf{X}^T\mathbf{Y} = \mathbf{X}\mathbf{B}\mathbf{M}\mathbf{R}^T\mathbf{M}^T\mathbf{B}^T\mathbf{X}^T\mathbf{Y}$ and that $\mathbf{R}^T = \mathbf{R}$ we here have

$$\begin{aligned}
\left(\mathbf{B}\mathbf{M}\mathbf{R}\mathbf{M}^T\mathbf{B}^T\mathbf{X}^T\mathbf{X}\mathbf{P}\mathbf{L} - \mathbf{P}\mathbf{L}\right)^T\mathbf{X}^T\mathbf{Y} &= \mathbf{L}^T\mathbf{P}^T\mathbf{X}^T\mathbf{X}\mathbf{B}\mathbf{M}\mathbf{R}^T\mathbf{M}^T\mathbf{B}^T\mathbf{X}^T\mathbf{Y} - \mathbf{L}^T\mathbf{P}^T\mathbf{X}^T\mathbf{Y} \\
&= \mathbf{L}^T\mathbf{P}^T\left(\mathbf{X}^T\hat{\mathbf{Y}} - \mathbf{X}^T\mathbf{Y}\right) \\
&= \mathbf{L}^T\left(\mathbf{P}^T\mathbf{X}^T\mathbf{X}\mathbf{P}\left(\mathbf{P}^T\mathbf{X}^T\mathbf{X}\mathbf{P}\right)^{-1}\mathbf{P}^T\mathbf{X}^T\mathbf{Y} - \mathbf{P}^T\mathbf{X}^T\mathbf{Y}\right) \\
&= \mathbf{L}^T\left(\mathbf{P}^T\mathbf{X}^T\mathbf{Y} - \mathbf{P}^T\mathbf{X}^T\mathbf{Y}\right) = \mathbf{0}, \qquad (15)
\end{aligned}$$

i.e. also in the general case we have

$$\tilde{\mathbf{B}} = \mathbf{B}\mathbf{M}\left(\mathbf{M}^T\mathbf{B}^T\mathbf{X}^T\mathbf{X}\mathbf{B}\mathbf{M}\right)^{-1}\mathbf{M}^T\mathbf{B}^T\mathbf{X}^T\mathbf{Y} = \mathbf{B} = \hat{\mathbf{B}}. \qquad (16)$$

References

[1] Svensson O, Kourti T, MacGregor JF. An investigation of orthogonal signal correction algorithms and their characteristics. *J. Chemometrics* 2002; **16**: 176-188.

[2] Trygg J. *Parsimonious Multivariate Models.* PhD thesis, Umeå University, 2002.

[3] Kvalheim OM, Karstang T. Interpretation of Latent-Variable Regression Models, *Chemometrics and Intelligent Laboratory Systems* 1989; **7**: 39-51.

[4] Ergon R. Compression into two-component PLS factorization. *J. Chemometrics* 2003; **17**: 303-312.

[5] Ergon R. Informative PLS score-loading plots for process understanding and monitoring. *J. Process Control* 2004; **14**: 889-897.

[6] Ergon R. Informative Score-Loading Plots for Multi-Response Process Monitoring. In Pomeransev AL (Ed.) *Progress in Chemometrics Research*, Nova Science Publishers, New York, 2005.

[7] Martens H, Næs T. *Multivariate Calibration.* Wiley: New York, 1989.

[8] Höskuldsson A. *Prediction Methods in Science and Technology, Vol. 1 Basic Theory.* Thor Publishing: Copenhagen, 1996.

[9] Hodouin D, MacGregor JF, Hou M, Franklin M. Multivariate Statistical Analysis of Mineral Processing Plant Data. *Can. Inst. Mining Bull. 86* 1993; No. 975, 23-34.

[10] Ergon R. PLS score-loading correspondence and a bi-orthogonal factorization. *J. Chemometrics* 2002; **16**: 368-373.

[11] Kailath T. *Linear Systems.* Prentice-Hall: Englewood Cliffs, New Jersey, 1980.