# Finding $\mathbf{Y}$-relevant part of $\mathbf{X}$ by use of PCR and PLSR model reduction methods

Rolf Ergon
rolf.ergon@hit.no
Telemark University College, Porsgrunn, Norway

Published in Journal of Chemometrics 2007; **21**: 537-546

**Abstract**

The paper is considering the following question: Using principal component regression (PCR) or partial least squares regression (PLSR), how much data can be removed from $\mathbf{X}$ while retaining the original ability to predict $\mathbf{Y}$? Two model reduction methods using similarity transformations are discussed, one giving projections of original loadings onto the column space of the fitted response matrix $\hat{\mathbf{Y}}$ (essentially the orthogonal signal correction (OSC) methods), and one giving projections of original scores onto the column space of the coefficient matrix $\hat{\mathbf{B}}$ (essentially the net analyte signal (NAS) methods). The loading projection method gives model residuals that are orthogonal to $\mathbf{Y}$ and $\hat{\mathbf{Y}}$, which is valuable in certain applications. The score projection method, on the other hand, gives model residuals that are orthogonal to $\hat{\mathbf{B}}$, which is essential in other applications. It is shown that the reduced matrix $\mathbf{X}_\mathbf{Y}^\mathrm{S}$ from the score projection method is a subset of the reduced matrix $\mathbf{X}_\mathbf{Y}^\mathrm{L}$ from the loading projection method. It therefore has the smallest Frobenius norm, and thus the smallest total column variance, assuming centered data.

KEYWORDS: PCR/PLSR model reduction; similarity transformations; OPLS, NAS; minimum $\mathbf{Y}$-relevant part; Frobenius norm

## 1 Introduction

Principal component regression and partial least squares regression (PCR and PLSR) are well known methods for solution of ill-posed multivariate regression problems. Both methods make use of factorizations of the regressor and response data matrices into $\mathbf{X} = \sum_{i=1}^{A} \mathbf{t}_i \mathbf{p}_i^T + \mathbf{E} = \hat{\mathbf{X}} + \mathbf{E}$ and $\mathbf{Y} = \sum_{i=1}^{A} \mathbf{t}_i \mathbf{q}_i^T + \mathbf{F} = \hat{\mathbf{Y}} + \mathbf{F}$, where the number of components $A$ with score vectors $\mathbf{t}_i$ and loading vectors $\mathbf{p}_i$ and $\mathbf{q}_i$ is determined through either cross-validation or test set validation, and where $\mathbf{E}$ and $\mathbf{F}$ are unmodeled residuals. The number of components in such latent variables (LV) models are often higher than strictly necessary, and methods for model reduction are therefore of interest. One reason for this is that interpretations of score and loading plots are easier with fewer components, as discussed in references given below, and in an industrial data example in Subsection 4. The present paper will, however, primarily focus on a different aspect, as illustrated in Fig. 1:

- How much data can be removed from $\mathbf{X}$, without loss of the original ability to predict $\mathbf{Y}$ ?
- In other words, what is the truly smallest possible $\mathbf{Y}$-relevant part $\mathbf{X}_\mathbf{Y}$ of $\mathbf{X}$ ?

As a measure of the size of $\mathbf{X}_\mathbf{Y}$ we may use the Frobenius norm, defined in Section 3 below. This will also be a measure of the total column variance of $\mathbf{X}_\mathbf{Y}$, assuming centered data.

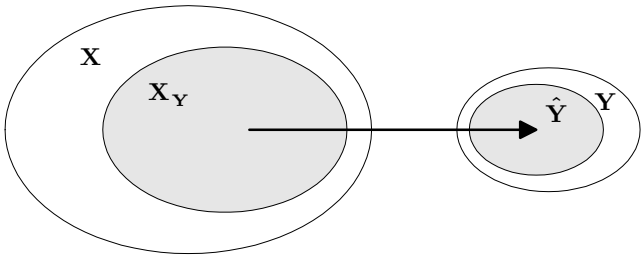


Figure 1. Illustration of data matrices $\mathbf{Y}$ and $\mathbf{X}$, with $\mathbf{Y}$-relevant part $\mathbf{X}_\mathbf{Y}$.

One method for model reduction is to identify and remove the $\mathbf{Y}$-orthogonal part of $\mathbf{X}$, which is the aim of the preprocessing orthogonal signal correction (OSC) methods [1,2], e.g. the OPLS algorithm [3]. The $\mathbf{Y}$-orthogonal part of $\mathbf{X}$ can also be found by a post-processing similarity transformation of the original PCA/PCR or PLS factorization [4]. The starting point in reference [4] was the non-orthogonalized PLS factorization [5], where in the single response case and as illustrated in Fig. 2, all score vectors except the first one are orthogonal to the fitted response vector $\hat{\mathbf{y}}$. Hence, the similarity transformation only has to split $\mathbf{t}_1$ into one component $\mathbf{t}_1^{\mathrm{ST}}$ in the direction of $\hat{\mathbf{y}}$ and one component orthogonal to $\hat{\mathbf{y}}$, while the score vectors in $\mathbf{T}_{2:A}$ (columns 2 to $A$ of $\mathbf{T}$) should be left as they are. It was also shown in Reference [4] that the results of this within a second similarity transformation are identical with the results from a slightly modified version of the OPLS algorithm (OPLS with non-orthogonalized PLS).

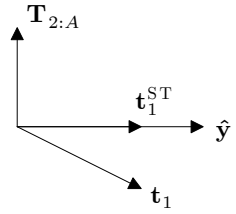


Figure 2. Score vectors in relation to $\hat{\mathbf{y}}$ for non-orthogonalized PLS factorization of $\mathbf{X}$. Here, $\mathbf{T}_{2:A}$ stands for columns 2 to $A$ of the non-orthogonalized score matrix $\mathbf{T}$.

For the orthogonalized PLS factorization [5], the situation is different. As illustrated in Fig. 3, all the orthogonal score vectors must here be split into components in the direction of and orthogonal to $\hat{\mathbf{y}}$, but that can also be done with a similarity transformation (see Section 2).

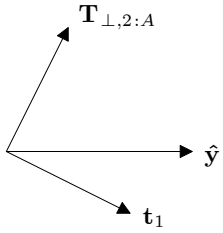


Figure 3. Score vectors in relation to $\hat{\mathbf{y}}$ for orthogonalized PLS factorization of $\mathbf{X}$. Here, $\mathbf{T}_{\perp,2:A}$ stands for columns 2 to $A$ of the orthogonalized score matrix $\mathbf{T}_\perp$.

As shown in Section 2 the post-processing similarity transformation method illustrated in Figures 2 and 3 can be extended to cover also multi-response cases, with a response matrix $\mathbf{Y}$ and a fitted matrix $\hat{\mathbf{Y}}$. The common effect of all these similarity transformations is that the original loadings in the space spanned by the score vectors are projected onto the column space of $\hat{\mathbf{Y}}$. We will therefore refer to these methods as loading projection methods (although an alternative reference could have been score vector projections).

An alternative reduction method is obviously to project the original scores onto the column space of the coefficient matrix $\hat{\mathbf{B}}$, and by doing so we can identify and remove $\hat{\mathbf{B}}$-orthogonal parts of $\mathbf{X}^T$. Such a

projection was suggested already in Reference [6], and it has been used in definitions of net analytic signal (NAS) [7,8,9]. An example related to this is the 2PLS algorithm presented in Reference [10] and intended for process monitoring applications, where the projection subplane includes $\hat{\mathbf{b}}$, and a more general treatment is given in Reference [11]. As illustrated in Fig. 4, all loading vectors must here be split into components in the direction of and orthogonal to $\hat{\mathbf{B}}$, which as shown in Section 2 can also be done with a similarity transformation. We will refer to this type of reduction methods as score projection methods (although an alternative reference could here have been loading vector projections).

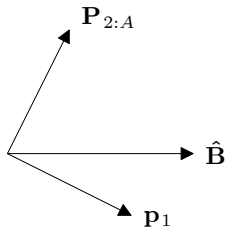


Figure 4. Loading vectors in relation to $\hat{\mathbf{B}}$ for a general LV factorization of $\mathbf{X}$. Here, $\mathbf{P}_{2:A}$ stands for columns 2 to $A$ of the loading matrix $\mathbf{P}$.

The fact that model reduction can be obtained through either loading or score projections onto reduced subspaces (removing $\mathbf{Y}$- and $\hat{\mathbf{Y}}$-orthogonal parts from $\mathbf{X}$ or $\hat{\mathbf{B}}$-orthogonal parts from $\mathbf{X}^T$), raises interesting questions:

- Will the two approaches give identical results? As shown in Section 2, the answer is *no*.
- Which method will give the smallest $\mathbf{Y}$-relevant part of $\mathbf{X}$, in some reasonable sense? As shown in Section 3, $\mathbf{X}_{\mathbf{Y}}^{\mathrm{S}}$ from the score projection method is a subset of $\mathbf{X}_{\mathbf{Y}}^{\mathrm{L}}$ from the loading projection method, and it therefore has the smallest Frobenius norm, and thus the smallest total row and column variance. In that respect it is a better method for this purpose than use of OSC methods.
- Does the reduction result in a standalone reduced LV model? As shown in Section 3, the answer is *yes* for the score projection method, and *no* for the loading projection method.

# 2 Model reduction by similarity transformations

## 2.1 Latent variables model

Let us in the following use the non-orthogonalized PLSR factorization as an example. Results for PCR follow in corresponding and straightforward ways, while results for orthogonalized PLSR are summarized in remarks below. As a starting point we use the LV model

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^T + \mathbf{F} \tag{1}$$
$$\mathbf{X} = \mathbf{T}\mathbf{W}^T + \mathbf{E}, \tag{2}$$

where we assume $m$ independent responses and $A \geq m$ components, and where $\mathbf{F}$ and $\mathbf{E}$ are unmodeled residuals. We thus have $\hat{\mathbf{Y}} = \mathbf{T}\mathbf{Q}^T$, where $\mathbf{Q}^T$ is found from the least squares (LS) solution $\mathbf{Q}^T = \left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{Y}$. The loading weights matrix $\mathbf{W}$ is orthonormal, and from the LS solution $\mathbf{T} = \mathbf{X}\mathbf{W}\left(\mathbf{W}^T\mathbf{W}\right)^{-1} = \mathbf{X}\mathbf{W}$ thus follows $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{W}\mathbf{Q}^T$, i.e. the regression coefficients

$$\hat{\mathbf{B}} = \mathbf{W}\mathbf{Q}^T = \mathbf{W}\left(\mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W}\right)^{-1}\mathbf{W}^T\mathbf{X}^T\mathbf{Y}. \tag{3}$$

Here, $\mathbf{W}$ is found by use of the NIPALS PLSR algorithm [5].

**Remark 1** *It is straightforward to show that Eq. (3) is valid also if* $\mathbf{W}$ *is not orthonormal, i.e. for all LS regressions of* $\mathbf{Y}$ *on* $\mathbf{T} = \mathbf{X}\mathbf{W}$ *for any* $\mathbf{W}$.

For what follows it is important to note that $\mathbf{EW} = \mathbf{0}$, and thus also $\mathbf{E}\hat{\mathbf{B}} = \mathbf{EWQ^T} = \mathbf{0}$, while on the other hand $\mathbf{T}^T\mathbf{E} \neq \mathbf{0}$ (see Appendix A for proofs of these and some other orthogonality properties). For simplicity of presentation we will assume that $\mathbf{Y}$ has full rank, otherwise it should be replaced by an appropriate number of principal components. With $m$ independent responses, we will also need at least $A = m$ components in order to obtain good predictions of all responses.

## 2.2 Loading projection transformation

Introducing an invertible transformation matrix $\mathbf{M}_\mathrm{L}$ the LV model (1,2) gives

$$\mathbf{Y} = \mathbf{T}\mathbf{M}_\mathrm{L}\mathbf{M}_\mathrm{L}^{-1}\mathbf{Q}^T + \mathbf{F} = \tilde{\mathbf{T}}_\mathrm{L}\tilde{\mathbf{Q}}_\mathrm{L}^T + \mathbf{F} \tag{4}$$

$$\mathbf{X} = \mathbf{T}\mathbf{M}_\mathrm{L}\mathbf{M}_\mathrm{L}^{-1}\mathbf{W}^T + \mathbf{E} = \tilde{\mathbf{T}}_\mathrm{L}\tilde{\mathbf{W}}_\mathrm{L}^T + \mathbf{E}. \tag{5}$$

Under the given assumptions we have $A \geq m$ components, and it is then straightforward to show that (using the notation $\mathbf{Q} = \left[\begin{array}{cc} \mathbf{Q}_{1:m} & \mathbf{Q}_{m+1:A} \end{array}\right]$ etc.)

$$\mathbf{M}_\mathrm{L} = \left[\begin{array}{cc} \mathbf{Q}_{1:m}^T & -\left(\hat{\mathbf{Y}}^T\mathbf{T}_{1:m}\right)^{-1}\hat{\mathbf{Y}}^T\mathbf{T}_{m+1:A} \\ \mathbf{Q}_{m+1:A}^T & \mathbf{I} \end{array}\right] \tag{6}$$

gives

$$\tilde{\mathbf{T}}_{\mathrm{L},1:m} = \mathbf{T}\mathbf{Q}^T = \hat{\mathbf{Y}}, \tag{7}$$

while

$$\tilde{\mathbf{T}}_{\mathrm{L},m+1:A} = -\mathbf{T}_{1:m}\left(\hat{\mathbf{Y}}^T\mathbf{T}_{1:m}\right)^{-1}\hat{\mathbf{Y}}^T\mathbf{T}_{m+1:A} + \mathbf{T}_{m+1:A} \tag{8}$$

is orthogonal to $\hat{\mathbf{Y}}$.

From $\hat{\mathbf{Y}} = \tilde{\mathbf{T}}_\mathrm{L}\tilde{\mathbf{Q}}_\mathrm{L}^T$ and $\tilde{\mathbf{T}}_{\mathrm{L},1:m} = \hat{\mathbf{Y}}$ according to Eqs. (4) and (7) follows $\tilde{\mathbf{Q}}_\mathrm{L} = \left[\begin{array}{cc} \mathbf{I} & \mathbf{0} \end{array}\right]$, and from Eq. (5) and the fact that $\tilde{\mathbf{T}}_{\mathrm{L},1:m}^T\tilde{\mathbf{T}}_{\mathrm{L},m+1:A} = \hat{\mathbf{Y}}^T\tilde{\mathbf{T}}_{\mathrm{L},m+1:A} = \mathbf{0}$ follows the LS solution

$$\tilde{\mathbf{W}}_\mathrm{L}^T = \left(\tilde{\mathbf{T}}_\mathrm{L}^T\tilde{\mathbf{T}}_\mathrm{L}\right)^{-1}\tilde{\mathbf{T}}_\mathrm{L}^T\mathbf{T}\mathbf{W}^T = \left[\begin{array}{c} \left(\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}\right)^{-1}\hat{\mathbf{Y}}^T\mathbf{T}\mathbf{W}^T \\ \left(\tilde{\mathbf{T}}_{\mathrm{L},m+1:A}^T\tilde{\mathbf{T}}_{\mathrm{L},m+1:A}\right)^{-1}\tilde{\mathbf{T}}_{\mathrm{L},m+1:A}^T\mathbf{T}\mathbf{W}^T \end{array}\right]. \tag{9}$$

Note that we here cannot replace $\mathbf{TW}^T$ with $\mathbf{X}$, for the reason that $\mathbf{T}$ and thus $\tilde{\mathbf{T}}_\mathrm{L} = \mathbf{T}\mathbf{M}_\mathrm{L}$ are not orthogonal to the residual $\mathbf{E}$. The results for $\tilde{\mathbf{Q}}_\mathrm{L}$ and $\tilde{\mathbf{W}}_\mathrm{L}$ may with some effort also be obtained from $\tilde{\mathbf{Q}}_\mathrm{L}^T = \mathbf{M}_\mathrm{L}^{-1}\mathbf{Q}^T$ and $\tilde{\mathbf{W}}_\mathrm{L}^T = \mathbf{M}_\mathrm{L}^{-1}\mathbf{W}^T$.

In summary, the structured information in $\mathbf{X}$ is split into two parts resulting in

$$\mathbf{X} = \hat{\mathbf{Y}}\left(\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}\right)^{-1}\hat{\mathbf{Y}}^T\mathbf{T}\mathbf{W}^T + \tilde{\mathbf{T}}_{\mathrm{L},m+1:A}\tilde{\mathbf{W}}_{\mathrm{L},m+1:A}^T + \mathbf{E}, \tag{10}$$

where $\tilde{\mathbf{T}}_{\mathrm{L},m+1:A}$ is orthogonal to $\hat{\mathbf{Y}}$ (and $\tilde{\mathbf{W}}_{\mathrm{L},m+1:A}^T$ is orthogonal to $\hat{\mathbf{B}}$). Note that the second block column in $\mathbf{M}_\mathrm{L}$ may be multiplied from the right by any invertible matrix, resulting in different similarity transformations of $\tilde{\mathbf{T}}_{\mathrm{L},m+1:A}\tilde{\mathbf{W}}_{\mathrm{L},m+1:A}^T$, but not affecting $\hat{\mathbf{Y}}\left(\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}\right)^{-1}\hat{\mathbf{Y}}^T\mathbf{T}\mathbf{W}^T$.

**Remark 2** *For orthogonalized PLSR [5] using the factorization $\mathbf{X} = \mathbf{T}_\perp\mathbf{P}^T + \mathbf{E}_\perp$ (where $\mathbf{E}_\perp$ is somewhat different from $\mathbf{E}$ in Eq. (2)) the result corresponding to Eq. (10) is obtained by replacing $\mathbf{W}^T$ with $\mathbf{P}^T$, or by replacing $\mathbf{TW}^T$ with $\mathbf{X}$ (since $\mathbf{T}_\perp$ is orthogonal to $\mathbf{E}_\perp$). For the single response case, the first $\mathbf{y}$-relevant part will then be exactly the same as with use of the OPLS algorithm [3], while the $\mathbf{y}$-orthogonal parts will be identical within a similarity transformation (different $\tilde{\mathbf{T}}_\perp$ and $\tilde{\mathbf{P}}$, but the same product $\tilde{\mathbf{T}}_\perp\tilde{\mathbf{P}}^T$, see also related results in [4]).*

**Remark 3** *For orthogonalized PLSR the loading matrix of the $\mathbf{Y}$-relevant part is $\mathbf{X}^T\hat{\mathbf{Y}}\left(\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}\right)^{-1}$, which should be compared with the ordinary LS result $\mathbf{X}^T\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}$ for spectrum profile estimation (see also Reference [12] for a discussion).*

### 2.3 Score projection transformations

As a starting point we here use the LV model (1,2) and an invertible transformation matrix $\mathbf{M}_{\mathrm{S}}$, giving

$$\mathbf{Y} = \mathbf{T}\mathbf{M}_{\mathrm{S}}^{-T}\mathbf{M}_{\mathrm{S}}^{T}\mathbf{Q}^{T} + \mathbf{F} = \tilde{\mathbf{T}}_{\mathrm{S}}\tilde{\mathbf{Q}}_{\mathrm{S}}^{T} + \mathbf{F} \tag{11}$$

$$\mathbf{X} = \mathbf{T}\mathbf{M}_{\mathrm{S}}^{-T}\mathbf{M}_{\mathrm{S}}^{T}\mathbf{W}^{T} + \mathbf{E} = \tilde{\mathbf{T}}_{\mathrm{S}}\tilde{\mathbf{W}}_{\mathrm{S}}^{T} + \mathbf{E}. \tag{12}$$

It is now straightforward to show that

$$\mathbf{M}_{\mathrm{S}} = \begin{bmatrix} \mathbf{Q}_{1:m}^{T} & -\left(\hat{\mathbf{B}}^{T}\mathbf{W}_{1:m}\right)^{-1}\hat{\mathbf{B}}^{T}\mathbf{W}_{m+1:A} \\ \mathbf{Q}_{m+1:A}^{T} & \mathbf{I} \end{bmatrix} \tag{13}$$

gives

$$\tilde{\mathbf{W}}_{\mathrm{S},1:m} = \mathbf{W}\mathbf{Q}^{T} = \hat{\mathbf{B}}, \tag{14}$$

while

$$\tilde{\mathbf{W}}_{\mathrm{S},m+1:A} = -\mathbf{W}_{1:m}\left(\hat{\mathbf{B}}^{T}\mathbf{W}_{1:m}\right)^{-1}\hat{\mathbf{B}}^{T}\mathbf{W}_{m+1:A} + \mathbf{W}_{m+1:A} \tag{15}$$

is orthogonal to $\hat{\mathbf{B}}$.

From Eq. (12) and the fact that $\tilde{\mathbf{W}}_{\mathrm{S},1:m}^{T}\tilde{\mathbf{W}}_{\mathrm{S},m+1:A} = \hat{\mathbf{B}}^{T}\tilde{\mathbf{W}}_{\mathrm{S},m+1:A} = \mathbf{0}$ follows the LS solution

$$\tilde{\mathbf{T}}_{\mathrm{S}} = \mathbf{X}\tilde{\mathbf{W}}_{\mathrm{S}}\left(\tilde{\mathbf{W}}_{\mathrm{S}}^{T}\tilde{\mathbf{W}}_{\mathrm{S}}\right)^{-1} = \begin{bmatrix} \hat{\mathbf{Y}}\left(\hat{\mathbf{B}}^{T}\hat{\mathbf{B}}\right)^{-1} & \mathbf{X}\tilde{\mathbf{W}}_{\mathrm{S},m+1:A}\left(\tilde{\mathbf{W}}_{\mathrm{S},m+1:A}^{T}\tilde{\mathbf{W}}_{\mathrm{S},m+1:A}\right)^{-1} \end{bmatrix}, \tag{16}$$

where we make use of the fact that $\mathbf{E}\hat{\mathbf{B}} = \mathbf{0}$. From $\hat{\mathbf{Y}} = \tilde{\mathbf{T}}_{\mathrm{S}}\tilde{\mathbf{Q}}_{\mathrm{S}}^{T}$ thus also follows $\tilde{\mathbf{Q}}_{\mathrm{S}} = \begin{bmatrix} \hat{\mathbf{B}}^{T}\hat{\mathbf{B}} & \mathbf{0} \end{bmatrix}$. The results for $\tilde{\mathbf{Q}}_{\mathrm{S}}$ and $\tilde{\mathbf{T}}_{\mathrm{S}}$ may also be obtained by use of $\mathbf{M}_{\mathrm{S}}^{-1}$.

In summary, the structured information in $\mathbf{X}$ is now split into two parts resulting in

$$\mathbf{X} = \hat{\mathbf{Y}}\left(\hat{\mathbf{B}}^{T}\hat{\mathbf{B}}\right)^{-1}\hat{\mathbf{B}}^{T} + \tilde{\mathbf{T}}_{\mathrm{S},m+1:A}\tilde{\mathbf{W}}_{\mathrm{S},m+1:A}^{T} + \mathbf{E}, \tag{17}$$

where $\tilde{\mathbf{W}}_{\mathrm{S},m+1:A}$ is orthogonal to $\hat{\mathbf{B}}$ (while $\tilde{\mathbf{T}}_{\mathrm{S},m+1:A}$ is not orthogonal to $\hat{\mathbf{Y}}$). Also here the second block column of $\mathbf{M}_{\mathrm{S}}$ may be multiplied from the right by any invertible matrix, with a similarity transformation of $\tilde{\mathbf{T}}_{\mathrm{S},m+1:A}\tilde{\mathbf{W}}_{\mathrm{S},m+1:A}^{T}$ as result.

**Remark 4** *For orthogonalized PLSR using the LV model* $\mathbf{Y} = \mathbf{T}_{\perp}\mathbf{Q}_{\perp}^{T} + \mathbf{F}$ *and* $\mathbf{X} = \mathbf{T}_{\perp}\mathbf{P}^{T} + \mathbf{E}_{\perp}$ *(where* $\mathbf{E}_{\perp} \neq \mathbf{E}$ *), a factorization corresponding to Eq. (17) cannot be obtained. The reason for this is that the columns of* $\hat{\mathbf{B}}$ *are found in the column space of* $\mathbf{W}$ *and not of* $\mathbf{P}$*. This is an argument for using the factorization* $\mathbf{X} = \mathbf{T}_{\perp}\mathbf{P}^{T}\mathbf{W}\mathbf{W}^{T} + \mathbf{E}$*, where* $\mathbf{T}_{\perp}\mathbf{P}^{T}\mathbf{W}$ *is equal to* $\mathbf{T}$ *in Eq. (2), as also argued for in Reference [4]. Also using* $\mathbf{X} = \mathbf{T}_{\perp}\mathbf{P}^{T} + \mathbf{E}_{\perp}$*, however, we can construct the first* $\mathbf{Y}$*-relevant part of Eq. (17) as soon as* $\hat{\mathbf{B}}$ *is determined. Note here that* $\hat{\mathbf{B}}$ *is the same as for non-orthogonalized PLSR.*

**Remark 5** *The* $\mathbf{Y}$*-relevant part of Eq. (17) applied to a new sample, i.e.* $\mathbf{x}_{new}^{S} = \hat{\mathbf{B}}\left(\hat{\mathbf{B}}^{T}\hat{\mathbf{B}}\right)^{-1}\hat{\mathbf{B}}^{T}\mathbf{x}_{new}$*, is a multiresponse generalization of NAS according to the definition "The NAS vector is the part of the mixture spectrum that is useful for prediction" [8].*

## 3 Comparison of the two model reduction methods

### 3.1 General comparison

As shown in subsections below, the basic difference between the methods discussed above is that the first $\mathbf{Y}$-relevant part on the right hand side of Eq. (17), $\mathbf{X}_{\mathbf{Y}}^{\mathrm{S}} = \hat{\mathbf{Y}}\left(\hat{\mathbf{B}}^{T}\hat{\mathbf{B}}\right)^{-1}\hat{\mathbf{B}}^{T}$, is a subset of the first $\mathbf{Y}$-relevant part on the right hand side of Eq. (10), $\mathbf{X}_{\mathbf{Y}}^{\mathrm{L}} = \hat{\mathbf{Y}}\left(\hat{\mathbf{Y}}^{T}\hat{\mathbf{Y}}\right)^{-1}\hat{\mathbf{Y}}^{T}\mathbf{T}\mathbf{W}^{T}$ (see Theorem 1 with proof below). The score projection method will thus remove all $\mathbf{Y}$-orthogonal information from the modeled part of $\mathbf{X}$, just as the loading projection method will do (this is the main objective of the OSC/OPLS methods). But in addition it will remove some other information that is not necessary for prediction of $\mathbf{Y}$.

In some more detail the following general differences should also be noted:

- The loading projection method isolates all information related to $\mathbf{Y}$ in the first part $\mathbf{X}_{\mathbf{Y}}^{\mathrm{L}}$, leaving $\mathbf{Y}$-orthogonal information only in the residual second part. This may certainly be beneficial in some applications, while the drawback in other applications may be that the loading matrices in the two parts are not orthogonal.
- The score projection method isolates as little information as possible in the $\mathbf{Y}$-relevant part $\mathbf{X}_{\mathbf{Y}}^{\mathrm{S}}$, leaving not only $\mathbf{Y}$-orthogonal information in the residual second part. In this case, however, the loading matrices in the two parts are orthogonal, and this is a useful property in some applications (see process monitoring example in Section 5 below).
- The loading projection residuals may be used for analysis of $\mathbf{Y}$-orthogonal structured information, while the score projection residuals may be used for analysis of $\hat{\mathbf{B}}$-orthogonal structured information.
- An additional difference is that the score projection method results in a standalone reduced model, which is not the case for the loading projection method (see discussion below).

A thorough application oriented comparison of the two methods is beyond the aim of the present theoretical paper, and many applications related to chemical, biological, genetic etc. data are presumably not yet developed. However, a single process monitoring example in Section 4 below will illustrate the usefulness of the score projection method.

### 3.2 Relation between the two Y-relevant parts

We focus here on the first $\mathbf{Y}$-relevant parts on the right hand sides of Eqs. (10) and (17), $\mathbf{X}_{\mathbf{Y}}^{\mathrm{L}} = \hat{\mathbf{Y}}\left(\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}\right)^{-1}\hat{\mathbf{Y}}^T\mathbf{T}\mathbf{W}^T$ and $\mathbf{X}_{\mathbf{Y}}^{\mathrm{S}} = \hat{\mathbf{Y}}\left(\hat{\mathbf{B}}^T\hat{\mathbf{B}}\right)^{-1}\hat{\mathbf{B}}^T$. The relation between these parts are given by the following theorem and illustrated in Fig. 5, and as a result Fig. 1 may be altered into Fig. 6:

**Theorem 1** *Using $\hat{\mathbf{Y}}$ as common score matrix for $X_{\mathbf{Y}}^{\mathrm{S}}$ and $X_{\mathbf{Y}}^{\mathrm{L}}$, the loading matrix of $X_{\mathbf{Y}}^{\mathrm{S}}$ is obtained by projection of the loading vectors of $X_{\mathbf{Y}}^{\mathrm{L}}$ onto the column space of $\hat{\mathbf{B}}$. For the special case of $A = m$, i.e. for as many original components as the number of responses, the two loading matrices are equal.*

**Proof.** Projection of the column vectors in the loading matrix $\mathbf{W}\mathbf{T}^T\hat{\mathbf{Y}}\left(\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}\right)^{-1}$ of $\mathbf{X}_{\mathbf{Y}}^{\mathrm{L}}$ (using $\hat{\mathbf{Y}}$ as score matrix) onto the column space of $\hat{\mathbf{B}}$, results in $\hat{\mathbf{B}}\left(\hat{\mathbf{B}}^T\hat{\mathbf{B}}\right)^{-1}\hat{\mathbf{B}}^T\mathbf{W}\mathbf{T}^T\hat{\mathbf{Y}}\left(\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}\right)^{-1}$ $= \hat{\mathbf{B}}\left(\hat{\mathbf{B}}^T\hat{\mathbf{B}}\right)^{-1}\mathbf{Q}\mathbf{W}^T\mathbf{W}\mathbf{T}^T\mathbf{T}\mathbf{Q}^T\left(\mathbf{Q}\mathbf{T}^T\mathbf{T}\mathbf{Q}^T\right)^{-1} = \hat{\mathbf{B}}\left(\hat{\mathbf{B}}^T\hat{\mathbf{B}}\right)^{-1}$, where the relations $\hat{\mathbf{Y}} = \mathbf{T}\mathbf{Q}^T$ and $\hat{\mathbf{B}} = \mathbf{W}\mathbf{Q}^T$ from Eqs. (1) and (3) and the fact that $\mathbf{W}^T\mathbf{W} = \mathbf{I}$ come to use. The fact that the projection results in the loading matrix of $\mathbf{X}_{\mathbf{Y}}^{\mathrm{S}}$ (again using $\hat{\mathbf{Y}}$ as score matrix), shows that $\mathbf{X}_{\mathbf{Y}}^{\mathrm{L}} = \mathbf{X}_{\mathbf{Y}}^{\mathrm{S}} + \hat{\mathbf{Y}}\mathbf{Z}^T$, where $\mathbf{Z}$ is orthogonal to $\hat{\mathbf{B}}$. For the special case of $A = m$ the matrix $\mathbf{Q}$ is invertible, such that $\mathbf{X}_{\mathbf{Y}}^{\mathrm{L}} = \mathbf{T}\mathbf{Q}^T\left(\mathbf{Q}\mathbf{T}^T\mathbf{T}\mathbf{Q}^T\right)^{-1}\mathbf{Q}\mathbf{T}^T\mathbf{T}\mathbf{W}^T = \mathbf{T}\mathbf{W}^T$, while $\mathbf{X}_{\mathbf{Y}}^{\mathrm{S}} = \mathbf{T}\mathbf{Q}^T\left(\mathbf{Q}\mathbf{W}^T\mathbf{W}\mathbf{Q}^T\right)^{-1}\mathbf{Q}\mathbf{W}^T = \mathbf{T}\mathbf{W}^T$, which means that $\mathbf{Z} = \mathbf{0}$ and $\mathbf{X}_{\mathbf{Y}}^{\mathrm{L}} = \mathbf{X}_{\mathbf{Y}}^{\mathrm{S}}$. ■

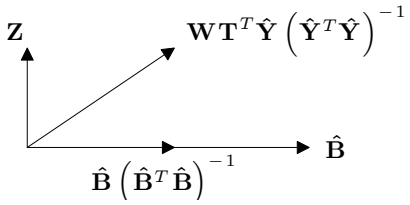


Figure 5. Relation between loading vectors of $\mathbf{X}_{\mathbf{Y}}^{\mathrm{L}}$ and $\mathbf{X}_{\mathbf{Y}}^{\mathrm{S}}$ (using $\hat{\mathbf{Y}}$ as score matrix).

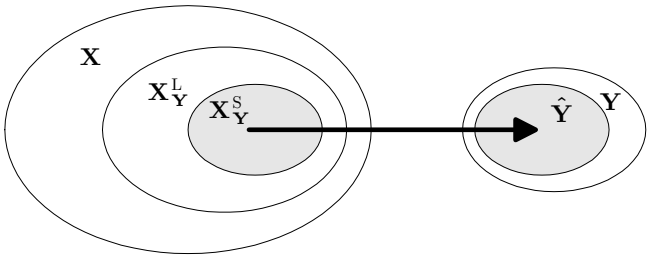


Figure 6. Modified illustration of data matrices $\mathbf{Y}$ and $\mathbf{X}$, with $\mathbf{Y}$-relevant parts $\mathbf{X}_\mathbf{Y}^\mathrm{L}$ and $\mathbf{X}_\mathbf{Y}^\mathrm{S}$.

### 3.3 Frobenius norms

The Frobenius norm of a matrix $\mathbf{X} \in \mathbb{R}^{N \times p}$ is defined as [13]

$$\|\mathbf{X}\|_\mathrm{F} = \sqrt{trace\left[\mathbf{X}\mathbf{X}^T\right]} = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{p} x_{ij}^2}, \tag{18}$$

i.e. as the square root of $N-1$ times the total column variance of $\mathbf{X}$, assuming centered data. For the score projection factorization (17) follows the Frobenius norm

$$\left\|\mathbf{X}_\mathbf{Y}^\mathrm{S}\right\|_\mathrm{F} = \sqrt{tr\left[\hat{\mathbf{Y}}\left(\hat{\mathbf{B}}^T\hat{\mathbf{B}}\right)^{-1}\hat{\mathbf{B}}^T\hat{\mathbf{B}}\left(\hat{\mathbf{B}}^\mathbf{T}\hat{\mathbf{B}}\right)^{-\mathbf{1}}\hat{\mathbf{Y}}^T\right]} = \sqrt{tr\left[\hat{\mathbf{Y}}\left(\hat{\mathbf{B}}^T\hat{\mathbf{B}}\right)^{-1}\hat{\mathbf{Y}}^T\right]}. \tag{19}$$

For the loading projection factorization (10), on the other hand, Theorem 1 with proof results in

$$\begin{aligned}\left\|\mathbf{X}_\mathbf{Y}^\mathrm{L}\right\|_\mathrm{F} &= \sqrt{tr\left[\hat{\mathbf{Y}}\left(\left(\hat{\mathbf{B}}^T\hat{\mathbf{B}}\right)^{-1}\hat{\mathbf{B}}^T + \mathbf{Z}^T\right)\left(\hat{\mathbf{B}}\left(\hat{\mathbf{B}}^T\hat{\mathbf{B}}\right)^{-1} + \mathbf{Z}\right)\hat{\mathbf{Y}}^T\right]} \\ &= \sqrt{tr\left[\hat{\mathbf{Y}}\left(\left(\hat{\mathbf{B}}^T\hat{\mathbf{B}}\right)^{-1} + \mathbf{Z}^T\mathbf{Z}\right)\hat{\mathbf{Y}}^T\right]} = \sqrt{tr\left[\hat{\mathbf{Y}}\left(\hat{\mathbf{B}}^T\hat{\mathbf{B}}\right)^{-1}\hat{\mathbf{Y}}^T\right] + tr\left[\hat{\mathbf{Y}}\mathbf{Z}^\mathbf{T}\mathbf{Z}\hat{\mathbf{Y}}^T\right]}.\end{aligned} \tag{20}$$

Since $trace\left[\hat{\mathbf{Y}}\mathbf{Z}^\mathbf{T}\mathbf{Z}\hat{\mathbf{Y}}^T\right]$ is positive, this shows that

$$\left\|\mathbf{X}_\mathbf{Y}^\mathrm{S}\right\|_\mathrm{F} \leq \left\|\mathbf{X}_\mathbf{Y}^\mathrm{L}\right\|_\mathrm{F}. \tag{21}$$

Equality is obtained for $A = m$, where $\mathbf{Z} = \mathbf{0}$.

### 3.4 Reduced models and prediction properties

The score projection factorization (17) forms the basis for a reduced model

$$\mathbf{Y} = \mathbf{T}_\mathrm{S}\mathbf{Q}_\mathrm{S}^T + \mathbf{F} \tag{22}$$

$$\mathbf{X} = \mathbf{T}_\mathrm{S}\mathbf{W}_\mathrm{S}^T + \mathbf{E}_\mathrm{S}, \tag{23}$$

where the loading matrix $\mathbf{W}_\mathrm{S} = \hat{\mathbf{B}}\left(\hat{\mathbf{B}}^T\hat{\mathbf{B}}\right)^{-\frac{1}{2}}$ is orthonormal (just as $\mathbf{W}$ in Eq. (2)), and where $\mathbf{Q}_\mathrm{S} = \left(\hat{\mathbf{B}}^T\hat{\mathbf{B}}\right)^{\frac{1}{2}}$ and $\mathbf{T}_\mathrm{S} = \hat{\mathbf{Y}}\left(\hat{\mathbf{B}}^T\hat{\mathbf{B}}\right)^{-\frac{1}{2}} = \mathbf{X}\mathbf{W}_\mathrm{S}$ (just as the score matrix in Eq. (2) is $\mathbf{T} = \mathbf{X}\mathbf{W}$). The reduced model (22,23) thus has the same score-loading correspondence properties as the original model (1,2) [14], which have been found useful in process monitoring methods [11,15,16], and it results in the same PLSR predictions (or PCR predictions, if a PCR model is used as a starting point). The regression coefficients $\hat{\mathbf{B}}$ may be found from the formula (3), using $\mathbf{W}_\mathrm{S}$ instead of $\mathbf{W}$, and a new sample will thus give the predictions $\hat{\mathbf{y}}_\mathrm{new}^T = \mathbf{x}_\mathrm{new}^T\mathbf{W}\left(\mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W}\right)^{-1}\mathbf{W}^T\mathbf{X}^T\mathbf{Y} = \mathbf{x}_\mathrm{new}^T\mathbf{W}_\mathrm{S}\left(\mathbf{W}_\mathrm{S}^T\mathbf{X}^T\mathbf{X}\mathbf{W}_\mathrm{S}\right)^{-1}\mathbf{W}_\mathrm{S}^T\mathbf{X}^T\mathbf{Y}$. If all of $\mathbf{X}$ except for $\mathbf{X}_\mathbf{Y}^\mathrm{S} =$

$\mathbf{T}_\mathrm{S}\mathbf{W}_\mathrm{S}^T$ is discarded, the model (22,23) is still valid (with zero residual), although a new PLSR computation using $\mathbf{X}_\mathbf{Y}^\mathrm{S}$ as input will result in a new loading weights matrix $\tilde{\mathbf{W}}_\mathrm{S}$. The coefficients $\hat{\mathbf{B}}$ and thus the predictions will still be the same, however, now computed as $\hat{\mathbf{y}}_\mathrm{new}^T = \mathbf{x}_\mathrm{new}^T\tilde{\mathbf{W}}_\mathrm{S}\left(\tilde{\mathbf{W}}_\mathrm{S}^T\left(\mathbf{X}_\mathbf{Y}^\mathrm{S}\right)^T\mathbf{X}_\mathbf{Y}^\mathrm{S}\tilde{\mathbf{W}}_\mathrm{S}\right)^{-1}\tilde{\mathbf{W}}_\mathrm{S}^T\left(\mathbf{X}_\mathbf{Y}^\mathrm{S}\right)^T\mathbf{Y}$. In these respects the score projection method results in a standalone reduced model.

The loading projection factorization (10), on the other hand, will not form the basis for a standalone reduced model. This is reflected in the fact that a new sample $\mathbf{x}_\mathrm{new}^T$ must be pretreated by removal of the $\mathbf{Y}$-orthogonal part according to Eq. (10), before the reduced model is used for prediction [3,4]. In order to do that we must make use of $\tilde{\mathbf{W}}_{\mathrm{L},m+1:A}^T$ in the $\mathbf{Y}$-orthogonal part of $\mathbf{X}$.

# 4 Industrial and laboratory data examples

## 4.1 Data sets

Three multiresponse data sets are used as examples, with all data centered and standardized:

- The Wentzell group at Dalhousie University has provided a data set under the name *gasoil* (http://www.dal.ca/~pdwentze/download.htm). The $\mathbf{X}$ data are UV spectra over 572 channels, and the number of response variables is four. The first 40 samples are here used for modeling, and samples 71-110 for validation.

- A data set originating from a mineral processing plant is published in Reference [17] (the *cleaner* data, originally published in Reference [18]). The problem considered here is to predict two given responses $y_4$ and $y_7$ from twelve known process variables. The first 40 samples are here used for modeling, and samples 181-220 for validation.

- The Cargill company and Eigenvector Research Inc. have provided a data set labeled *corn* (http://software.eigenvector.com/Data/Corn/index.html). From these data 80 samples of corn measured on a NIR spectrometer labeled m5 are used. 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. The first 40 samples are here used for modeling, and samples 41-80 for validation.

## 4.2 Comparison of multiresponse models

Table 1 summarizes root mean square error of prediction (RMSEP) and Frobenius norm results for the loading and score projection factorizations (10) and (17). The following procedure was followed for each of the data sets:

- The optimal number of original PLSR components, and the corresponding RMSEP values, were first determined by use of the NIPALS algorithm with the modeling data $\mathbf{X}$ and $\mathbf{Y}$ as inputs [5]. The original number of components $A$ and the resulting matrix of coefficients $\hat{\mathbf{B}}$ were noted.

- The loading projection factorization (10) was performed by determination of the transformation matrix $\mathbf{M}_\mathrm{L}$. The $\mathbf{Y}$-relevant first part of $\mathbf{X}$, $\mathbf{X}_\mathbf{Y}^\mathrm{L} = \hat{\mathbf{Y}}\left(\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}\right)^{-1}\hat{\mathbf{Y}}^T\mathbf{T}\mathbf{W}^T$, was then together with $\mathbf{Y}$ used in a new PLSR computation with as many components as number of $\mathbf{Y}$ variables, and the resulting coefficient matrix $\hat{\mathbf{B}}_\mathrm{L}$ was noted. The validation data $\mathbf{X}_\mathrm{val}$ was pretreated according to Eq. (10), i.e. $\mathbf{X}_\mathrm{val}^\mathrm{red} = \mathbf{X}_\mathrm{val} - \tilde{\mathbf{T}}_{\mathrm{L},m+1:A}^\mathrm{val}\tilde{\mathbf{P}}_{\mathrm{L},1+m:A}^T = \mathbf{X}_\mathrm{val} - \mathbf{X}_\mathrm{val}\mathbf{W}\mathbf{M}_{\mathrm{L},2}\tilde{\mathbf{P}}_{\mathrm{L},1+m:A}^T$, where $\mathbf{M}_{\mathrm{L},2}$ is the second column of $\mathbf{M}_\mathrm{L}$. Finally, the RMSEP values were determined by use of the prediction error $\mathbf{Y}_\mathrm{val} - \mathbf{X}_\mathrm{val}^\mathrm{red}\hat{\mathbf{B}}_\mathrm{L}$.

- The score projection factorization (17) was performed by determination of the transformation matrix $\mathbf{M}_\mathrm{S}$. The $\mathbf{Y}$-relevant first part of $\mathbf{X}$, $\mathbf{X}_\mathbf{Y}^\mathrm{S} = \hat{\mathbf{Y}}\left(\hat{\mathbf{B}}^T\hat{\mathbf{B}}\right)^{-1}\hat{\mathbf{B}}^T$, was then together with $\mathbf{Y}$ used in a new PLSR computation with as many components as number of $\mathbf{Y}$ variables, and the resulting coefficient matrix $\hat{\mathbf{B}}_\mathrm{S}$ was noted. Finally, the RMSEP value were determined by use of the prediction error $\mathbf{Y}_\mathrm{val} - \mathbf{X}_\mathrm{val}\hat{\mathbf{B}}_\mathrm{S}$.

- Finally, the Frobenius norms in Table 1 were determined.

Table 1: Various Frobenius norms for three data sets, based on 40 modeling samples and non-orthogonalized multiresponse PLSR (PLS2). RMSEP values were obtained by use of 40 test set samples. For each data set, the RMSEP values determined as described in the text for the original model and the two reduced models were identical.

| | Gasoil data | Cleaner data | Corn data |
|---|---|---|---|
| Number of variables | 572 | 12 | 700 |
| Number of PLSR components for original model | 6 | 6 | 15 |
| Number of responses | 4 | 2 | 4 |
| Number of PLSR components for reduced models | 4 | 2 | 4 |
| RMSEP for response 1 | 0.1331 | 0.2030 | 0.3561 |
| RMSEP for response 2 | 0.1643 | 0.3490 | 0.9004 |
| RMSEP for response 3 | 0.1985 | | 0.4332 |
| RMSEP for response 4 | 0.3160 | | 0.4783 |
| $\left\|\mathbf{X}\right\|_{\mathrm{F}}$ | 145.6108 | 18.5925 | 165.2266 |
| $\left\|\mathbf{X}_{\mathbf{Y}}^{\mathrm{L}}\right\|_{\mathrm{F}}$ | 121.0159 | 12.5143 | 144.0545 |
| $\left\|\mathbf{X}_{\mathbf{Y}}^{\mathrm{S}}\right\|_{\mathrm{F}}$ | 64.6440 | 10.2373 | 2.9796 |
| $\left\|\left(\hat{\mathbf{B}}^T\hat{\mathbf{B}}\right)^{-1}\hat{\mathbf{B}}\right\|_{\mathrm{F}}$ | 486.0401 | 1.6603 | 0.6325 |
| $\left\|\hat{\mathbf{Y}}\right\|_{\mathrm{F}}$ | 12.3876 | 8.7445 | 12.3835 |

Note that $\left\|\mathbf{X}_{\mathbf{Y}}^{\mathrm{S}}\right\|_{\mathrm{F}} < \left\|\mathbf{X}_{\mathbf{Y}}^{\mathrm{L}}\right\|_{\mathrm{F}}$ for all data sets. Also note the very similar $\left\|\hat{\mathbf{Y}}\right\|_{\mathrm{F}}$ and large difference in $\left\|\mathbf{X}_{\mathbf{Y}}^{\mathrm{S}}\right\|_{\mathrm{F}}$ for the *gasoil* and *corn* data. This is due to the fact that the numerical values in $\hat{\mathbf{B}}_{\mathrm{gasoil}}$ generally are much smaller ($\sqrt{tr(\hat{\mathbf{B}}_{\mathrm{gasoil}}^T\hat{\mathbf{B}}_{\mathrm{gasoil}})^{-1}} = 486$) than the values in $\hat{\mathbf{B}}_{\mathrm{corn}}$ ($\sqrt{tr(\hat{\mathbf{B}}_{\mathrm{corn}}^T\hat{\mathbf{B}}_{\mathrm{corn}})^{-1}} = 0.63$), although the column mean values are very similar.

When the residual $\mathbf{E}$ was added to the reduced matrices $\mathbf{X}_{\mathbf{Y}}^{\mathrm{L}}$ and $\mathbf{X}_{\mathbf{Y}}^{\mathrm{S}}$, i.e. when the second terms only in Eqs. (10) and (17) were removed, the RMSEP values were not the same as for the original model, but they were very similar. The reason is that the influences from the unstructured noise in $\mathbf{E}$ are different after removal of the second terms.

### 4.3 Process monitoring involving residual analysis

As shown in Section 3 above the score projection method gives the $\mathbf{Y}$-relevant part of $\mathbf{X}$ with the smallest Frobenius norm. From this also follows that it gives the largest residual after removal of the $\mathbf{Y}$-relevant part. These facts may potentially be utilized in many different application areas, and as an example we here use process monitoring.

Model reduction by use of the score projection method has been found useful for monitoring of processes with two response variables [16,17]. In such cases the natural choice is to project the scores onto the plane spanned by the two vectors of regression coefficients, $\hat{\mathbf{b}}_1$ and $\hat{\mathbf{b}}_2$. With one response variable only, the projection plane must in addition to $\hat{\mathbf{b}}$ be spanned by some other appropriate vector $\mathbf{v}$ in the space spanned by $\mathbf{W}$ (PLSR) or $\mathbf{P}$ (PCR). A natural choice of $\mathbf{v}$ is then the loading vector $\mathbf{p}_1$ of the first principal component of the residual of $\mathbf{X} - \mathbf{X}_{\mathbf{Y}}^{\mathrm{S}}$ (for added residual information and interpretation, we may also use score plots involving other residual components). The scores will then be plotted in the plane defined by the orthonormal loading matrix $\mathbf{P}_{\mathrm{plot}} = [\ \hat{\mathbf{b}}\left(\hat{\mathbf{b}}^T\hat{\mathbf{b}}\right)^{-0.5} \quad \mathbf{p}_1\ ]$, with a corresponding score matrix $\mathbf{T}_{\mathrm{plot}} = [\ \hat{\mathbf{y}}\left(\hat{\mathbf{b}}^T\hat{\mathbf{b}}\right)^{-0.5} \quad \left(\mathbf{X} - \mathbf{X}_{\mathbf{Y}}^{\mathrm{S}}\right)\mathbf{p}_1\ ]$. Assuming centered modeling data with $N$ samples, the score covariance

matrix is given by $\mathbf{S} = \frac{1}{N-1}\mathbf{T}_{\text{plot}}^T\mathbf{T}_{\text{plot}}$, and from this a confidence ellipse for the scores based on the upper control limit (UCL) for the Hotelling's $T^2$ statistics is computed from [21]

$$T_{\text{UCL}}^2 = \frac{2\left(N^2-1\right)}{N\left(N-2\right)}F_\alpha(2, N-1), \tag{24}$$

where $T^2$ for a given sample is given by $T_i^2 = \left[\begin{array}{cc} \hat{y}_i\left(\hat{\mathbf{b}}^T\hat{\mathbf{b}}\right)^{-0.5} & p_{i,1}\end{array}\right]\mathbf{S}^{-1}\left[\begin{array}{cc} \hat{y}_i\left(\hat{\mathbf{b}}^T\hat{\mathbf{b}}\right)^{-0.5} & p_{i,1}\end{array}\right]^T$. Since $\mathbf{P}_{\text{plot}}$ is orthonormal, there is total score-loading correspondence [14], and the contributions to a given score from the different variables can therefore be shown by contribution vectors in the score-loading biplot, as illustrated in Figure 7 below (where the first score vector $\hat{\mathbf{y}}\left(\hat{\mathbf{b}}^T\hat{\mathbf{b}}\right)^{-0.5}$ is scaled such that $\hat{\mathbf{y}}$ can be red directly from the axis). In order to indicate the direction of variable influences, the loadings are here plotted at equal distances from the origin.

As an example we use the Cleaner data presented above with $y_4$ as the single response variable, but for clarity of presentation we make use of the dominating $\mathbf{X}$ (in the projection used) variables number 2, 3, 4, 5, 8 and 10 only. As earlier the first 40 samples were used for PLSR modeling, now with $A = 3$ components, while samples 181-220 were used for testing, now with $RMSEP = 0.1757$. Figure 7 shows validation score number 191 approaching the UCL in a direction mainly orthogonal to the $\hat{y}$ axis. This is caused by positive values of variables 2 and 4 (attracting the score), and negative values of variables 3 and 5. If the score trace continues outside the confidence ellipse in that direction, the $\hat{y}$ value will still be close to target, but some process situation not represented in the modeling data would anyhow be indicated. Note that the sum of all six contribution vectors corresponds exactly to the score position, and that variables 6 and 8 has very little to say.

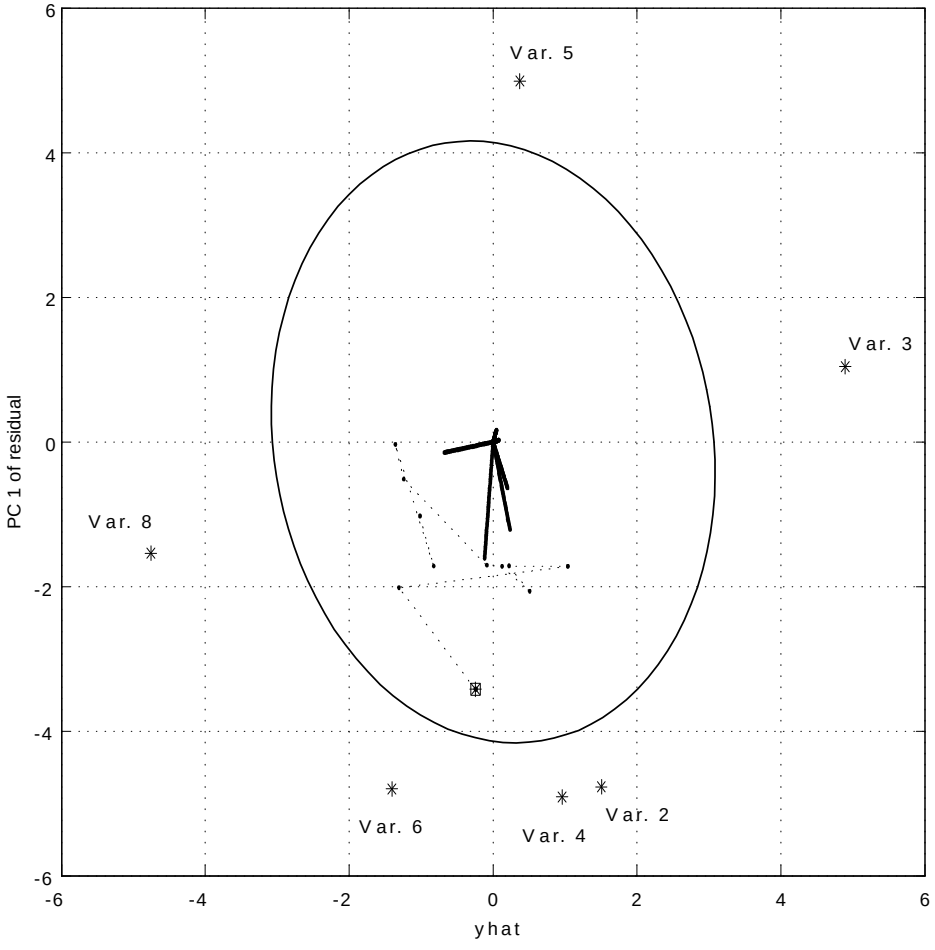

Figure 7. Score-loading-contribution plot with contribution vectors for sample 191, showing the score trace mainly moving in a direction orthogonal to the $\hat{y}$ axis (the dotted line shows score history). The $*$-markings with variable names are normalized loadings, showing the direction of variable influence on the scores.

Later, Fig. 8 shows score 209 falling slightly outside the confidence ellipse mainly in the direction of $\hat{y}$, indicating a potentially more serious failure situation. The contribution vectors show that this is caused by positive values of variables 2, 3 and 5, and a negative value of variable 4, while variables 6 and 8 also now has very little to say..

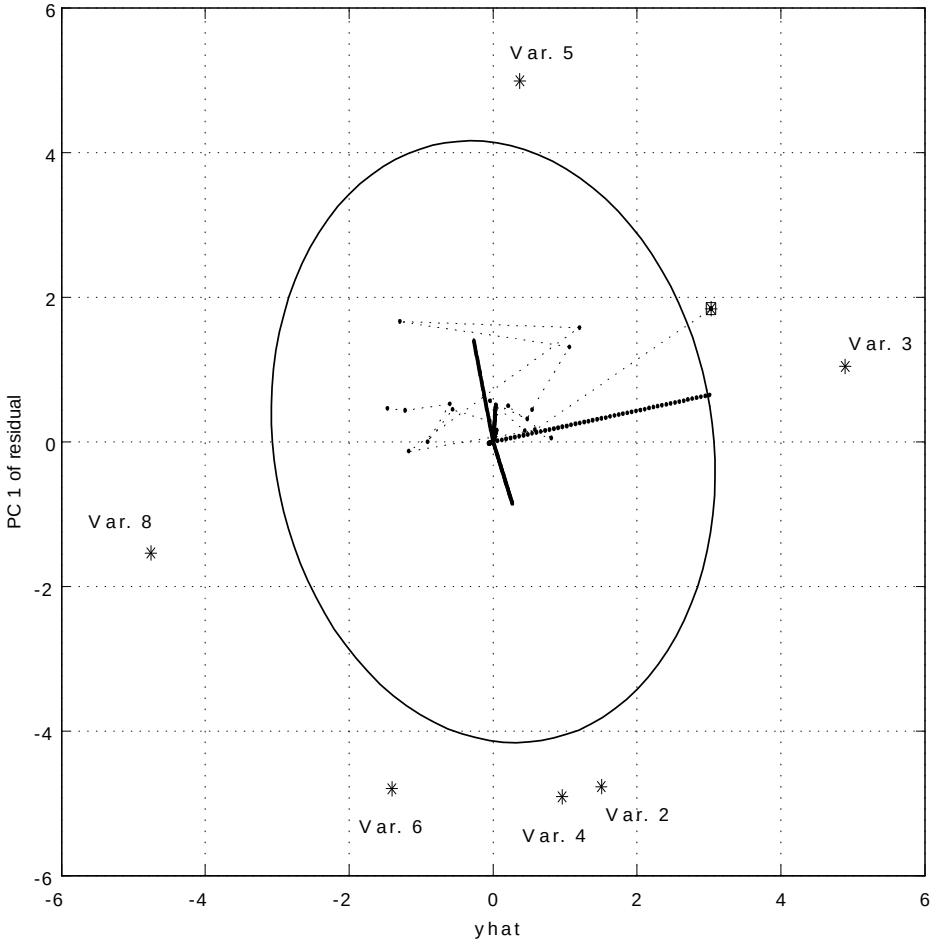


Figure 8. Score-loading-contribution plot for a sample 224, showing the score trace (dotted line) mainly moving in a direction of $\hat{y}$.

Note that the score projection application above is based on the fact that $\mathbf{X}_{\mathbf{Y}}^{\mathrm{S}}$ has $\hat{\mathbf{b}}\left(\hat{\mathbf{b}}^T\hat{\mathbf{b}}\right)^{-0.5}$ as loading vector, and that $\hat{\mathbf{b}}$ is orthogonal to the loading vectors of the residual. From this follows an orthonormal loading matrix $\mathbf{P}_{\mathrm{plot}}$, and thus exact score-loading correspondence [14]. The alternative use of the loading projection method would give a non-orthogonal matrix $\mathbf{P}_{\mathrm{plot}}^{\mathrm{L}} = \left[\begin{array}{cc} \mathbf{w}_1 & \mathbf{p}_1 \end{array}\right]$, and thus only approximate score-loading correspondence, depending on to which extent $\hat{\mathbf{b}}$ is dominated by $\mathbf{w}_1$.

# 5 Conclusion

In order to find the smallest part $\mathbf{X}_{\mathbf{Y}}^{\mathrm{S}}$ of $\mathbf{X}$ that can be used for explanation of $\mathbf{Y}$, one should remove all information in $\mathbf{X}^T$ orthogonal to $\hat{\mathbf{B}}$ using score projections (essentially as in the NAS methods). The result

is a reduced model (22,23) with the same basic properties as the original non-orthogonalized PLSR model (1,2), but with as many components as number of responses only.

The alternative use of loading projections (essentially as in the OSC/OPLS methods), where the goal is to remove information in $\mathbf{X}$ orthogonal to $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ (although also $\hat{\mathbf{B}}$-orthogonal parts of $\mathbf{X}^T$ may then be removed), isolates a part $\mathbf{X}_{\mathbf{Y}}^{\mathrm{L}}$ of $\mathbf{X}$. A comparison the two projection methods shows the following:

- The $\mathbf{Y}$-relevant part $\mathbf{X}_{\mathbf{Y}}^{\mathrm{S}}$ of the score projection factorization of $\mathbf{X}$ may also be obtained by a further projection of the corresponding part $\mathbf{X}_{\mathbf{Y}}^{\mathrm{L}}$ of the loading projection factorization. It thus has the smallest Frobenius norm and the smallest total column variance, assuming centered data. For the special case of as many original components as number of responses, the two norms are equal.
- The score projection method removes all $\mathbf{Y}$-orthogonal information from the modeled part of $\mathbf{X}$, just as the loading projection method does (this is the main objective of the OSC/OPLS methods). But in addition it removes some other information that is not necessary for prediction of $\mathbf{Y}$.
- The reduced score projection model is all that is needed for finding $\hat{\mathbf{B}}$ and thus for prediction of a new response $\mathbf{y}_{\mathrm{new}}$ from new regressor data $\mathbf{x}_{\mathrm{new}}$, and it may therefore be used as a standalone model.
- The score-loading correspondence property of the reduced score projection model makes it well suited for process monitoring applications, as shown in an example as well as in references.

The theoretical results including Theorem 1 on Frobenius norms, are substantiated by use of three industrial and laboratory data sets. The differences between $\left\|\mathbf{X}_{\mathbf{Y}}^{\mathrm{L}}\right\|_{\mathrm{F}}$ and $\left\|\mathbf{X}_{\mathbf{Y}}^{\mathrm{S}}\right\|_{\mathrm{F}}$ are clear, and in some cases quite large. Theorem 1 is so far a theoretical result only, and ideas about chemical, biological, genetic etc. data meaning and practical applications in addition to process monitoring remain to be developed.

# A Orthogonality properties of LV factorizations

We are considering here some orthogonality properties of the PLS factorizations $\mathbf{X} = \mathbf{T}_{\perp}\mathbf{P}^T + \mathbf{E}_{\perp}$ (orthogonalized) and $\mathbf{X} = \mathbf{T}\mathbf{W}^T + \mathbf{E}$ (non-orthogonalized). The following well established properties are assumed known:

- $\mathbf{T}_{\perp}^T\mathbf{T}_{\perp}$ is diagonal
- $\mathbf{T}^T\mathbf{T}$ is non-diagonal
- $\mathbf{W}^T\mathbf{W} = \mathbf{I}$
- $\mathbf{T} = \mathbf{X}\mathbf{W}$.

For the sake of completeness we also include orthogonality properties of the PCR factorization.

**Lemma 1** The product $\mathbf{P}^T\mathbf{W}$ has the bidiagonal structure

$$\mathbf{P}^T\mathbf{W} = \begin{bmatrix} 1 & \mathbf{p}_1^T\mathbf{w}_2 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1 & \mathbf{p}_{A-1}^T\mathbf{w}_A \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix}. \tag{25}$$

*Proof:* See Reference [19].

**Lemma 2** The relation between loading and loading weights vectors is

$$\mathbf{w}_{a+1} = \frac{\mathbf{w}_a - \mathbf{p}_a}{\|\mathbf{w}_a - \mathbf{p}_a\|} = \frac{\mathbf{w}_a - \mathbf{p}_a}{\sqrt{\mathbf{p}_a^T \mathbf{p}_a - 1}}. \tag{26}$$

*Proof:* This follows trivially from the NIPALS algorithm [6] and $\mathbf{W}^T\mathbf{W} = \mathbf{I}$.

**Lemma 3** The general vector product in $\mathbf{P}^T\mathbf{W}$ above is

$$\mathbf{p}_a^T \mathbf{w}_{a+1} = \frac{1 - \mathbf{p}_a^T \mathbf{p}_a}{\sqrt{\mathbf{p}_a^T \mathbf{p}_a - 1}} = -\sqrt{\mathbf{p}_a^T \mathbf{p}_a - 1}. \tag{27}$$

*Proof:* This follows directly from Lemma 1 and Lemma 2.

**Lemma 4** The factorizations $\mathbf{X} = \mathbf{T}_\perp \mathbf{W}^T \mathbf{W}\mathbf{W}^T + \mathbf{E}$ (revised orthogonalized) and $\mathbf{X} = \mathbf{T}\mathbf{W}^T + \mathbf{E}$ are identical.

*Proof:* From the two well known estimator expressions

$$\hat{\mathbf{B}} = \mathbf{W}\left(\mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W}\right)^{-1}\mathbf{W}^T\mathbf{X}^T\mathbf{Y} \tag{28}$$

and

$$\hat{\mathbf{B}} = \mathbf{W}\left(\mathbf{P}^T\mathbf{W}\right)^{-1}\mathbf{Q}_\perp^T = \mathbf{W}\left(\mathbf{P}^T\mathbf{W}\right)^{-1}\left(\mathbf{T}_\perp^T\mathbf{T}_\perp\right)^{-1}\mathbf{T}_\perp^T\mathbf{Y} \tag{29}$$

[20], follows

$$\mathbf{W}\left(\mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W}\right)^{-1}\mathbf{W}^T\mathbf{X}^T\mathbf{Y} = \mathbf{W}\left(\left(\mathbf{P}^T\mathbf{W}\right)^T\mathbf{T}_\perp^T\mathbf{T}_\perp\mathbf{P}^T\mathbf{W}\right)^{-1}\left(\mathbf{P}^T\mathbf{W}\right)^T\mathbf{T}_\perp^T\mathbf{Y},$$

i.e. $\mathbf{T}_\perp\mathbf{P}^T\mathbf{W} = \mathbf{X}\mathbf{W} = \mathbf{T}$.

**Lemma 5** The difference between the two residuals is

$$\mathbf{E}_\perp - \mathbf{E} = \mathbf{t}_{\perp,A}\left(\mathbf{w}_A^T - \mathbf{p}_A^T\right). \tag{30}$$

*Proof:* From the the revised orthogonalized factorization in Lemma 4 and Lemma 1 follows

$$\begin{aligned}
\mathbf{X} &= \mathbf{T}_\perp\mathbf{P}^T\mathbf{W}\mathbf{W}^T + \mathbf{E} = \begin{bmatrix} \mathbf{t}_{\perp,1} & \mathbf{t}_{\perp,2} & \cdots & \mathbf{t}_{\perp,A-1} & \mathbf{t}_{\perp,A} \end{bmatrix} \\
&\times \begin{bmatrix} 1 & \mathbf{p}_1^T\mathbf{w}_2 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1 & \mathbf{p}_{A-1}^T\mathbf{w}_A \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{w}_1^T \\ \mathbf{w}_2^T \\ \vdots \\ \mathbf{w}_{A-1}^T \\ \mathbf{w}_A^T \end{bmatrix} + \mathbf{E} \\
&= \mathbf{t}_{\perp,1}\left(\mathbf{w}_1^T + \mathbf{p}_1^T\mathbf{w}_2\mathbf{w}_2^T\right) + \cdots + \mathbf{t}_{\perp,A-1}\left(\mathbf{w}_{A-1}^T + \mathbf{p}_{A-1}^T\mathbf{w}_A\mathbf{w}_A^T\right) + \mathbf{t}_{\perp,A}\mathbf{w}_A^T + \mathbf{E},
\end{aligned} \tag{31}$$

and from Lemma 2, Lemma 3 and Lemma 4 thus follows

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T\mathbf{W}\mathbf{W}^T + \mathbf{E} = \mathbf{t}_{\perp,1}\mathbf{p}_1^T + \cdots + \mathbf{t}_{\perp,A-1}\mathbf{p}_{A-1}^T + \mathbf{t}_{\perp,A}\mathbf{w}_A^T + \mathbf{E}. \tag{32}$$

Comparison with the orthogonalized factorization

$$\mathbf{X} = \mathbf{T}_\perp\mathbf{P}^T + \mathbf{E}_\perp = \mathbf{t}_{\perp,1}\mathbf{p}_1^T + \cdots + \mathbf{t}_{\perp,A-1}\mathbf{p}_{A-1}^T + \mathbf{t}_{\perp,A}\mathbf{p}_A^T + \mathbf{E}_\perp \tag{33}$$

completes the proof.

**Property 1** The orthogonalized factorization has the property $\mathbf{T}_\perp^T\mathbf{E}_\perp = \mathbf{0}$.

*Proof:* Factorization with as many components as possible, i.e $A = p$, results in $p - A$ orthogonal score vectors in a complete factorization of $\mathbf{E}_\perp$. From this follows the property trivially.

**Property 2** The non-orthogonalized factorization has the property $\mathbf{EW} = \mathbf{0}$.
*Proof:* From $\mathbf{W}^T\mathbf{W} = \mathbf{I}$ and $\mathbf{T} = \mathbf{XW}$ follows $\mathbf{EW} = \left(\mathbf{X} - \mathbf{TW}^T\right)\mathbf{W} = \mathbf{T} - \mathbf{T} = \mathbf{0}$.

**Property 3** The orthogonalized factorization has the property $\mathbf{E}_\perp\mathbf{W} = \mathbf{0}$.
*Proof:* From $\mathbf{W}^T\mathbf{W} = \mathbf{I}$, Lemma 1, Lemma 5 and Property 2 follows

$$\begin{aligned}\mathbf{E}_\perp\mathbf{W} &= (\mathbf{E}_\perp - \mathbf{E})\mathbf{W} + \mathbf{EW} = \mathbf{t}_{\perp,A}\left(\mathbf{w}_A^T - \mathbf{p}_A^T\right)\begin{bmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \cdots & \mathbf{w}_A \end{bmatrix} \\ &= \mathbf{t}_{\perp,A}\left(\begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}\right) = \mathbf{0}.\end{aligned}$$

Note, however, that $\mathbf{E}_\perp\mathbf{P} \neq \mathbf{0}$ (not proved here).

**Property 4** The non-orthogonalized factorization has the property

$$\mathbf{T}^T\mathbf{E} = \mathbf{W}^T\mathbf{p}_A\mathbf{t}_{\perp,A}^T\mathbf{t}_{\perp,A}\left(\mathbf{p}_A^T - \mathbf{w}_A^T\right). \tag{34}$$

*Proof:* From $\mathbf{T} = \mathbf{XW}$, the orthogonality of $\mathbf{T}_\perp$, Lemma 1, Lemma 5 and Property 1 follows

$$\begin{aligned}\mathbf{T}^T\mathbf{E} &= \mathbf{W}^T\mathbf{P}\mathbf{T}_\perp^T\mathbf{E} = \mathbf{W}^T\mathbf{P}\mathbf{T}_\perp^T\left(\mathbf{E}_\perp - \mathbf{t}_{\perp,A}\left(\mathbf{w}_A^T - \mathbf{p}_A^T\right)\right) = \mathbf{W}^T\mathbf{P}\mathbf{T}_\perp^T\mathbf{t}_{\perp,A}\left(\mathbf{p}_A^T - \mathbf{w}_A^T\right) \\ &= \mathbf{W}^T\begin{bmatrix} \mathbf{p}_1 & \cdots & \mathbf{p}_A \end{bmatrix}\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \mathbf{t}_{\perp,A}^T\mathbf{t}_{\perp,A} \end{bmatrix}\left(\mathbf{p}_A^T - \mathbf{w}_A^T\right) = \mathbf{W}^T\mathbf{p}_A\mathbf{t}_{\perp,A}^T\mathbf{t}_{\perp,A}\left(\mathbf{p}_A^T - \mathbf{w}_A^T\right).\end{aligned}$$

Finally we include orthogonality properties of the PCR factorization $\mathbf{X} = \mathbf{T}_{\text{PCR}}\mathbf{P}_{\text{PCR}}^T + \mathbf{E}_{\text{PCR}}$.

**Property 5** The PCR factorization has the property $\mathbf{T}_{\text{PCR}}^T\mathbf{E}_{\text{PCR}} = \mathbf{0}$.
*Proof:* Using singular value decomposition (SVD) we obtain

$$\mathbf{X} = \begin{bmatrix} \mathbf{U} & \mathbf{U}_E \end{bmatrix}\begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_E \end{bmatrix}\begin{bmatrix} \mathbf{V}^T \\ \mathbf{V}_E^T \end{bmatrix} = \mathbf{USV}^T + \mathbf{U}_E\mathbf{S}_E\mathbf{V}_E^T = \mathbf{T}_{\text{PCR}}\mathbf{P}_{\text{PCR}}^T + \mathbf{E}_{\text{PCR}}.$$

Since $\mathbf{U}^T\mathbf{U}_E = \mathbf{0}$, this gives $\mathbf{T}_{\text{PCR}}^T\mathbf{E}_{\text{PCR}} = \mathbf{S}^T\mathbf{U}^T\mathbf{U}_E\mathbf{S}_E\mathbf{V}_E^T = \mathbf{0}$.

**Property 6** The PCR factorization has the property $\mathbf{E}_{\text{PCR}}\mathbf{P}_{\text{PCR}} = \mathbf{0}$.
*Proof:* Since the SVD above gives $\mathbf{V}_E^T\mathbf{V} = \mathbf{0}$ it also gives $\mathbf{E}_{\text{PCR}}\mathbf{P}_{\text{PCR}} = \mathbf{U}_E\mathbf{S}_E\mathbf{V}_E^T\mathbf{V} = \mathbf{0}$.

# References


[1] Wold S, Antti H, Lindgren F, Öhman J. Orthogonal signal correction of near-infrared spectra. *Chemometrics Intell. Lab. Syst.* 1998; **44**:175-185.

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

[3] Trygg J, Wold S. Orthogonal projections to latent structures. O-PLS. *J. Chemometrics* 2002; **16**: 119-128.

[4] Ergon R. PLS post-processing by similarity transformation (PLS+ST): a simple alternative to OPLS. *J. Chemometrics* 2005; **19**: 1-4.

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

[6] Kvalheim OM, Karstang T. Interpretation of Latent-Variable Regression Models. *Chemometrics Intell. Lab. Syst.* 1989; **7**: 39-51.

[7] Lorber A. Error Propagation and Figures of Merit for Quantification by Solving Matrix Equations. *Anal. Chem.* 1986; **58**: 1167-1172.

[8] Ferré J, Faber NM. Net analyte signal calculation for multivariate calibration. *Chemometrics Intell. Lab. Syst. 2003*; **69**: 123-136.

[9] Andersen CM, Bro R. Quantification and handling of sampling errors in instrumental measurements: a case study. *Chemometrics Intell. Lab. Syst. 2003*; **72**: 43-50.

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

[11] Ergon R. Reduced PCR/PLSR models by subspace projections. *Chemometrics Intell. Lab. Syst.* 2006; **81**: 68-73.

[12] Trygg J. Prediction and spectral profile estimation in multivariate calibration. *J. Chemometrics* 2004; **18**: 166-172.

[13] Golub G.H, Van Loan C.F. *Matrix Computations*. The Johns Hopkins University Press: Baltimore, 1996.

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

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

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

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

[18] 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.

[19] Manne R. Analysis of two partial-least-squares algorithms for multivariate calibration. *Chemometrics Intell. Lab. Syst.* 1987; **2**: 187-197.

[20] Helland IS. On the structure of partial least squares regression. *Commun. Statist.* 1988; **17**: 581-607.