Compression into two-component PLS factorizations

Rolf Ergon
Telemark University College
P.O.Box 203
N-3901 Porsgrunn, Norway
E-mail: Rolf.Ergon@hit.no
Telephone: ++ 47 35 57 51 60
Telefax: ++ 47 35 57 52 50

March 24, 2003

Abstract

Partial least squares regression (PLSR) often requires more than two components also in the case of a scalar response variable. As shown in papers on orthogonal signal correction (OSC) it is possible to reduce the number of components, resulting in easier data interpretation. In this paper it is shown how all scalar response PLSR models can be reduced to two-component models with the same structure and giving exactly the same estimator as the original model using many components. This is done by use of a direct and very simple algorithm, based on a two-dimensional subspace in the loading weights space. The resulting model may be transformed into different realizations for different purposes, e.g. latent variable profile estimation, process monitoring, fault detection etc., as discussed in the paper.

Keywords: PLS factorizations, parsimonious, model reduction

1 Introduction

Partial least squares regression (PLSR) is a well known and popular method for prediction of e.g. a scalar response variable y from multivariate regressor variables $\mathbf{x}^T$ according to $\hat{y}_{new} = \mathbf{x}_{new}^T \hat{\mathbf{b}}$ [e.g. 1]. Due to both a large number of regressor variables and collinearity, such regression problems are often ill-posed, and in PLSR a regularized estimator $\hat{\mathbf{b}}$ is found from modeling data collected in a regressor matrix $\mathbf{X}$ and a response vector $\mathbf{y}$. In many practical cases a fairly high number of PLS components are needed in order to obtain good predictions, and this makes pretreatment and interpretation of the data difficult. Reduction of the necessary number of PLS components is therefore a major advantage of orthogonal signal correction (OSC), which has attracted quite some attention in recent years [e.g. 2]. This raises the question whether it is possible to find a parsimonious PLS model in a more direct way than by use of OSC, and as shown in the present paper this can in fact be done, once the complete model with A components have been determined.

Reduction to one component only is trivial, this component is given by the estimator $\hat{\mathbf{b}}$ itself. The reason for use of more than one component is the interpretational advantages that score and loading plots give with respect to outlier detection, spectra estimation, process monitoring, fault detection and diagnosis etc. For the scalar response case this can be fully exploited when the

structured and $\mathbf{y}$-relevant variation in $\mathbf{X}$ is compressed into two components, and use of more components only complicates the situation.

Two components only may possibly be obtained in several ways other than OSC. One example is principal components of predictions (PCP) [3], where in the scalar response case $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}$ after normalization is used as one component, while residuals of $\mathbf{X}$ not contributing to $\hat{\mathbf{y}}$ are suggested for use as the second component.

In the present paper the goal is to develop an algorithm for a two-component PLS compression with the same structure and giving exactly the same estimator $\hat{\mathbf{b}}$ as the original PLS solution using many components. The basic insight behind this is illustrated in Fig. 1. The estimator $\hat{\mathbf{b}}$ is found in the space spanned by the loading weight vectors in $\hat{\mathbf{W}} = \begin{bmatrix} \hat{\mathbf{w}}_1 & \hat{\mathbf{w}}_2 & \cdots & \hat{\mathbf{w}}_A \end{bmatrix}$, i.e. it is a linear combination of these vectors. It is, however, also found in the plane defined by $\hat{\mathbf{w}}_1$ and a vector $\widetilde{\mathbf{w}}_2$, which is a linear combination of the vectors $\hat{\mathbf{w}}_2$, $\hat{\mathbf{w}}_3$, $\ldots$, $\hat{\mathbf{w}}_A$. The matrix $\tilde{\mathbf{W}} = \begin{bmatrix} \hat{\mathbf{w}}_1 & \widetilde{\mathbf{w}}_2 \end{bmatrix}$ is thus the loading weight matrix in a two-component PLS solution (2PLS) giving exactly the same estimator $\hat{\mathbf{b}}$ as the original solution using any number A components.

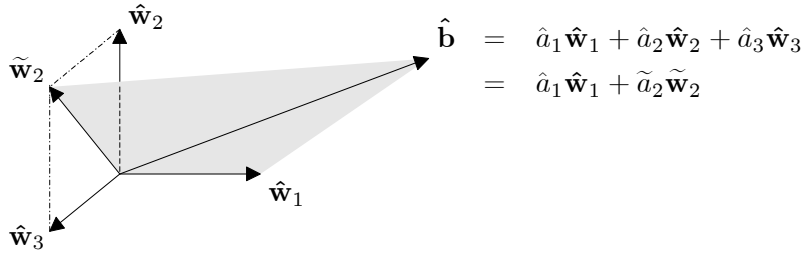

Figure 1. Illustration of basic insight behind the 2PLS factorization, assuming $A = 3$ original components. The PLSR estimator $\hat{\mathbf{b}}$ is found in the space spanned by $\hat{\mathbf{w}}_1$, $\hat{\mathbf{w}}_2$ and $\hat{\mathbf{w}}_3$, but also in the shadowed plane spanned by $\hat{\mathbf{w}}_1$ and $\widetilde{\mathbf{w}}_2$.

What matters in the original PLS model is not the matrix $\hat{\mathbf{W}}$ as such, but the space spanned by $\hat{\mathbf{w}}_1$, $\hat{\mathbf{w}}_2$, $\cdots$ $\hat{\mathbf{w}}_A$ [4], and in the 2PLS model it is the plane spanned by $\hat{\mathbf{w}}_1$ and $\widetilde{\mathbf{w}}_2$ that is essential. The reason for keeping $\hat{\mathbf{w}}_1$ will be apparent from the discussion on profile estimation in Section 3, where it is shown that it can be found from a well-posed least squares (LS) problem. Keeping $\hat{\mathbf{w}}_1$, there may be several meaningful ways to define the second loading vector for spanning the plane through $\hat{\mathbf{b}}$, and two options for different applications are presented in Section 2. Also note that all objects in $\mathbf{X}$ (row vectors) in the original PLS model are projected onto the space spanned by $\hat{\mathbf{w}}_2$, $\hat{\mathbf{w}}_3$, $\ldots$, $\hat{\mathbf{w}}_A$. They may thus be further projected onto the plane spanned by $\hat{\mathbf{w}}_1$ and $\widetilde{\mathbf{w}}_2$, and form a single score plot containing all $\mathbf{y}$-relevant information. If for some reason e.g. $\hat{\mathbf{w}}_2$ is more informative than $\hat{\mathbf{w}}_1$, a plane through $\hat{\mathbf{w}}_2$ and $\hat{\mathbf{b}}$ may be a better alternative. It will in any case result in a 2PLS model that gives the estimator $\hat{\mathbf{b}}$, as will in fact all planes through $\hat{\mathbf{b}}$ that are at the same time subspaces to the column space of $\hat{\mathbf{W}}$.

The reduction from many to only two PLS components should be helpful in a number of practical applications. A simulation example involving latent variable profile estimation as well as outlier detection is given in Section 4, and an industrial data example involving outlier detection and process monitoring for fault detection is given in Section 5. Another example is process monitoring by use of correspondence between score and loading plots [5], often presented in so-called biplots. With e.g. four PLS components it will then in principle be necessary to monitor three biplots in order to follow all possible process changes, while two components can be monitored by use of one biplot only.

The theory for compression into 2PLS factorizations is given in Section 2. The special case of latent variable profile estimation is discussed in Section 3, followed by a simulation example in Section 4. The industrial data example and final conclusions are given in Section 5 and Section 6. Some details are collected in Appendix A and B, while Matlab code is given in Appendix C.

2 Theory

2.1 PLS modeling

Multivariate calibration using PLSR directly or implicitly assumes a latent variables (LV) model

$$\begin{aligned} \mathbf{y} &= \mathbf{T}\mathbf{Q}^T + \mathbf{f} \quad (1) \\ \mathbf{X} &= \mathbf{T}\mathbf{L}^T + \mathbf{E}. \quad (2) \end{aligned}$$

The two PLSR algorithms of Wold and Martens [1] use different factorizations of $\mathbf{X}$ as summarized below, and thus also different factorizations of $\mathbf{y}$. They result, however, in the same estimator $\hat{\mathbf{b}}$, and they have the same first score vector $\hat{\mathbf{t}}_1$. The orthogonal scores PLSR algorithm of Wold is based on the factorization

$$\mathbf{X} = \hat{\mathbf{T}}_{\mathrm{W}}\hat{\mathbf{P}}^T\hat{\mathbf{W}}\hat{\mathbf{W}}^T + \hat{\mathbf{E}} = \hat{\mathbf{t}}_1\hat{\mathbf{p}}_1^T + \hat{\mathbf{t}}_2^{\mathrm{W}}\hat{\mathbf{p}}_2^T + \ldots + \hat{\mathbf{t}}_{A-1}^{\mathrm{W}}\hat{\mathbf{p}}_{A-1}^T + \hat{\mathbf{t}}_A^{\mathrm{W}}\hat{\mathbf{w}}_A^T + \hat{\mathbf{E}}, \quad (3)$$

where $\hat{\mathbf{T}}_{\mathrm{W}} = \mathbf{X}\hat{\mathbf{W}}\left(\hat{\mathbf{P}}^T\hat{\mathbf{W}}\right)^{-1}$ is orthogonal and $\hat{\mathbf{P}} = \begin{bmatrix} \hat{\mathbf{p}}_1 & \hat{\mathbf{p}}_2 & \cdots & \hat{\mathbf{p}}_A \end{bmatrix}$ is a special non-orthogonal loading matrix, while $\hat{\mathbf{W}} = \begin{bmatrix} \hat{\mathbf{w}}_1 & \hat{\mathbf{w}}_2 & \cdots & \hat{\mathbf{w}}_A \end{bmatrix}$ is the orthonormal loading weight matrix. The orthogonal loadings PLSR algorithm of Martens is based on the factorization

$$\mathbf{X} = \hat{\mathbf{T}}_{\mathrm{M}}\hat{\mathbf{W}}^T + \hat{\mathbf{E}} = \hat{\mathbf{t}}_1\hat{\mathbf{w}}_1^T + \hat{\mathbf{t}}_2^{\mathrm{M}}\hat{\mathbf{w}}_2^T + \ldots + \hat{\mathbf{t}}_A^{\mathrm{M}}\hat{\mathbf{w}}_A^T + \hat{\mathbf{E}}, \quad (4)$$

where $\hat{\mathbf{T}}_{\mathrm{M}} = \mathbf{X}\hat{\mathbf{W}}$ is non-orthogonal, while $\hat{\mathbf{W}}$ is the same as in the Wold algorithm. Note that $\hat{\mathbf{T}}_{\mathrm{W}}$ and $\hat{\mathbf{T}}_{\mathrm{M}}$ have the same column space. The estimator that is common for the two PLSR algorithms can be written as [4,6]

$$\hat{\mathbf{b}} = \hat{\mathbf{W}}\left(\hat{\mathbf{W}}^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{W}}\right)^{-1}\hat{\mathbf{W}}^T\mathbf{X}^T\mathbf{y}. \quad (5)$$

As a part of the PLSR algorithms the first loading weight vector is chosen as

$$\hat{\mathbf{w}}_1 = c_1\mathbf{X}^T\mathbf{y} = \left(\mathbf{y}^T\mathbf{X}\mathbf{X}^T\mathbf{y}\right)^{-\frac{1}{2}}\mathbf{X}^T\mathbf{y}. \quad (6)$$

A good argument for that is given in the discussion on LV profile estimation in Section 3.

2.2 Compression into two PLS components

As shown in Section 3 below the first loading weight vector $\hat{\mathbf{w}}_1$ can be found from a well-posed LS problem, and it gives valuable information about the latent variable corresponding to the response y. The other loading weight vectors $\hat{\mathbf{w}}_2$ to $\hat{\mathbf{w}}_A$ contain information about the other latent variables, but the interpretation is difficult due to confounding. Although more than two PLS components may be needed to obtain the best predictive performance, the added model complexity will thus give interpretational difficulties. This is the motivation for compression of the original model using A components into a final model using two components only.

As illustrated in Fig. 1, the central problem is to find a second vector that together with $\hat{\mathbf{w}}_1$ spans the shadowed plane that includes $\hat{\mathbf{b}}$. One way of doing this follows from the estimator formulation (5), in that $\left(\hat{\mathbf{W}}^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{W}}\right)^{-1}\hat{\mathbf{W}}^T\mathbf{X}^T\mathbf{y} = \begin{bmatrix} \hat{a}_2 & \hat{a}_3 & \cdots & \hat{a}_A \end{bmatrix}^T$, and thus

$$\hat{\mathbf{b}} = \hat{a}_1\hat{\mathbf{w}}_1 + \begin{bmatrix} \hat{\mathbf{w}}_2 & \hat{\mathbf{w}}_3 & \cdots & \hat{\mathbf{w}}_A \end{bmatrix}\begin{bmatrix} \hat{a}_2 & \hat{a}_3 & \cdots & \hat{a}_A \end{bmatrix}^T = \hat{a}_1\hat{\mathbf{w}}_1 + \widetilde{a}_2\widetilde{\mathbf{w}}_2. \quad (7)$$

We summarize the 2PLS compression and its properties in Theorem 1 below (see Appendix A for proof, and Appendix C for an algorithm in Matlab code). The second vector spanning the shadowed plane in Fig. 1 is not necessarily $\widetilde{\mathbf{w}}_2$, but it is convenient to make use of the orthogonal vectors $\hat{\mathbf{w}}_1$ and $\widetilde{\mathbf{w}}_2$ in the theorem.

Theorem 1 The original PLSR estimator (5) can be written as

$$\hat{\mathbf{b}} = \tilde{\mathbf{W}} \left(\tilde{\mathbf{W}}^T \mathbf{X}^T \mathbf{X} \tilde{\mathbf{W}} \right)^{-1} \tilde{\mathbf{W}}^T \mathbf{X}^T \mathbf{y}, \tag{8}$$

where $\tilde{\mathbf{W}} = \begin{bmatrix} \hat{\mathbf{w}}_1 & \widetilde{\mathbf{w}}_2 \end{bmatrix}$ is the new orthonormal loading (weight) matrix. The corresponding factorization of $\mathbf{X}$ is

$$\mathbf{X} = \tilde{\mathbf{T}}_{\mathrm{M}} \tilde{\mathbf{W}}^T + \tilde{\mathbf{E}} = \hat{\mathbf{t}}_1 \hat{\mathbf{w}}_1^T + \tilde{\mathbf{t}}_2^{\mathrm{M}} \widetilde{\mathbf{w}}_2^T + \tilde{\mathbf{E}}, \tag{9}$$

where $\widetilde{\mathbf{w}}_2$ is

$$\widetilde{\mathbf{w}}_2 = \frac{\begin{bmatrix} \hat{\mathbf{w}}_2 & \hat{\mathbf{w}}_3 & \cdots & \hat{\mathbf{w}}_A \end{bmatrix} \begin{bmatrix} \hat{a}_2 & \hat{a}_3 & \cdots & \hat{a}_A \end{bmatrix}^T}{\left\| \begin{bmatrix} \hat{\mathbf{w}}_2 & \hat{\mathbf{w}}_3 & \cdots & \hat{\mathbf{w}}_A \end{bmatrix} \begin{bmatrix} \hat{a}_2 & \hat{a}_3 & \cdots & \hat{a}_A \end{bmatrix}^T \right\|}. \tag{10}$$

Here $\begin{bmatrix} \hat{a}_2 & \hat{a}_3 & \cdots & \hat{a}_A \end{bmatrix}^T$ is extracted from $\begin{bmatrix} \hat{a}_1 & \hat{a}_2 & \cdots & \hat{a}_A \end{bmatrix}^T = \left(\hat{\mathbf{W}}^T \mathbf{X}^T \mathbf{X} \hat{\mathbf{W}} \right)^{-1} \hat{\mathbf{W}}^T \mathbf{X}^T \mathbf{y}$, while $\hat{\mathbf{t}}_1 = \mathbf{X}\hat{\mathbf{w}}_1$ is the same as in the factorizations (3) and (4), and $\tilde{\mathbf{t}}_2^{\mathbf{M}} = \mathbf{X}\widetilde{\mathbf{w}}_2$. Furthermore, $\tilde{\mathbf{t}}_2^{\mathbf{M}}$ is orthogonal to both $\mathbf{y}$ and $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}$, i.e. $\mathbf{y}^T \tilde{\mathbf{t}}_2^{\mathbf{M}} = 0$ and $\hat{\mathbf{y}}^T \tilde{\mathbf{t}}_2^{\mathbf{M}} = 0$.

■

Note that the residual $\tilde{\mathbf{E}}$ may be different from the original residual $\hat{\mathbf{E}}$, i.e. some extra $\mathbf{y}$-orthogonal structured variation in $\mathbf{X}$ may be captured in $\tilde{\mathbf{E}}$ (see Subsection 2.4 below, and industrial data example in Section 5).

Remark 1 *Since $\hat{\mathbf{W}}$ in the estimator (5) may be replaced by $\hat{\mathbf{W}}\mathbf{M}$, where $\mathbf{M}$ is any invertible transformation matrix, it follows from Theorem 1 that any plane containing $\hat{\mathbf{b}}$ that is also a subspace to the column space of $\hat{\mathbf{W}}$ may be used instead of the shadowed plane in Fig. 1. The theorem may thus be given a more general formulation.*

The factorization (9) has orthonormal loadings, just as the Martens factorization (4). It has the nice property that it gives exact correspondence between score and loading directions, i.e. an object $\mathbf{x}_i^T = \begin{bmatrix} 0 & \cdots & 0 & \Delta x_{ij} & 0 & \cdots & 0 \end{bmatrix}$ is found in the same direction in the score plot as variable number j in the loading plot. This property may be utilized in process monitoring, and for that purpose also a bi-orthogonal representation is possible, with orthogonal score and loading matrices [5] Process monitoring applications combining such score-loading correspondence with the 2PLS compression will be reported separately. A transformation to a form with orthogonal scores is also useful, and follows next.

2.3 Transformation to orthogonal scores form

In order to retain $\hat{\mathbf{w}}_1$ and the nice property that $\tilde{\mathbf{t}}_2^{\mathrm{M}}$ is orthogonal to both $\mathbf{y}$ and $\hat{\mathbf{y}}$, we write the factorization (9) as $\mathbf{X} = \tilde{\mathbf{T}}_{\mathrm{M}} \mathbf{M} \mathbf{M}^{-1} \tilde{\mathbf{W}}^T + \tilde{\mathbf{E}}$, and use the transformation matrix $\mathbf{M} = \begin{bmatrix} 1 & 0 \\ d & f \end{bmatrix}$. With $d = -\hat{\mathbf{t}}_1^T \tilde{\mathbf{t}}_2^{\mathrm{M}} / \left(\tilde{\mathbf{t}}_2^{\mathrm{M}} \right)^T \tilde{\mathbf{t}}_2^{\mathrm{M}}$ and $f = \sqrt{d^2 + 1}$ we obtain an orthogonal score matrix $\tilde{\mathbf{T}}_{\mathrm{W}}$ and a non-orthogonal loading matrix $\tilde{\mathbf{P}}_{\mathrm{W}}$, just as in the Wold factorization (6),

$$\mathbf{X} = \tilde{\mathbf{T}}_{\mathrm{W}} \tilde{\mathbf{P}}_{\mathrm{W}}^T + \tilde{\mathbf{E}} = \tilde{\mathbf{t}}_1^{\mathrm{W}} \hat{\mathbf{w}}_1^T + \tilde{\mathbf{t}}_2^{\mathrm{W}} \left(\tilde{\mathbf{p}}_2^{\mathrm{W}} \right)^T + \tilde{\mathbf{E}}, \tag{11}$$

(see Appendix B for proof). We also obtain $\left(\tilde{\mathbf{p}}_2^{\mathrm{W}} \right)^T \tilde{\mathbf{p}}_2^{\mathrm{W}} = \hat{\mathbf{w}}_1^T \hat{\mathbf{w}}_1 = 1$.

Since $\tilde{\mathbf{t}}_2^{\mathrm{W}}$ is orthogonal to both $\tilde{\mathbf{t}}_1^{\mathrm{W}}$ and $\hat{\mathbf{y}}$, and since $\hat{\mathbf{y}}$ is found in the plane spanned by $\tilde{\mathbf{t}}_1^{\mathrm{W}}$ and $\tilde{\mathbf{t}}_2^{\mathrm{W}}$, $\tilde{\mathbf{t}}_1^{\mathrm{W}}$ will have the same direction as $\hat{\mathbf{y}}$. Note that the scalar relation $\tilde{t}_1^{\mathrm{W}} = c\hat{y}$ holds also for new objects $\mathbf{x}_{new}^T$ (see Appendix B for proof). This may for example be utilized in process monitoring applications, where score movements in the $\tilde{\mathbf{t}}_1^{\mathrm{W}}$ direction would mean predicted response changes, while movements in the $\tilde{\mathbf{t}}_2^{\mathrm{W}}$ direction would mean $\mathbf{y}$-orthogonal process changes that are not captured in $\tilde{\mathbf{E}}$ (see industrial data example in Section 5). This is basically the same properties

as in score plots based PCP [3], except that a PCP score plot for a scalar response variable doesn't filter out $\mathbf{y}$-orthogonal process changes in the same way. A score plot based on the 2PLS algorithm shows only $\mathbf{y}$-orthogonal process changes that are reflected in $\tilde{\mathbf{X}} = \tilde{\mathbf{T}}_{\mathrm{W}}\tilde{\mathbf{P}}_{\mathrm{W}}^T$, while a PCP score plot shows all $\mathbf{y}$-orthogonal process changes contributing to the first principal component of the residual after removal of the first (and only) PCP component from $\mathbf{X}$ (see Section 5 for a comparison using industrial data).

2.4 Discussion on residuals

Assume an orthogonal loadings PLS factorization (4) with $A = 3$ components. A given object $\mathbf{x}^T$ may then be plotted in the coordinate system in Fig. 1, with projections (scores) $\hat{t}_1 = \mathbf{x}^T\hat{\mathbf{w}}_1$, $\hat{t}_2^{\mathrm{M}} = \mathbf{x}^T\hat{\mathbf{w}}_2$ and $\hat{t}_3^{\mathrm{M}} = \mathbf{x}^T\hat{\mathbf{w}}_3$ on the $\hat{\mathbf{w}}_1$, $\hat{\mathbf{w}}_2$ and $\hat{\mathbf{w}}_3$ axes, and accordingly included in $\hat{\mathbf{X}} = \hat{\mathbf{T}}_{\mathrm{M}}\hat{\mathbf{W}}^T$. Objects with $\hat{t}_1 = 0$ will fall on the plane spanned by $\hat{\mathbf{w}}_2$ and $\hat{\mathbf{w}}_3$, and objects with $\hat{t}_1 = \hat{t}_2^{\mathrm{M}} = \hat{t}_3^{\mathrm{M}} = 0$ will be totally captured in the residual $\hat{\mathbf{E}}$. Some of the objects that fall on the plane spanned by $\hat{\mathbf{w}}_2$ and $\hat{\mathbf{w}}_3$ may, however, be orthogonal to the new vector $\widetilde{\mathbf{w}}_2$, and thus not included in $\tilde{\mathbf{X}} = \tilde{\mathbf{T}}_{\mathrm{M}}\tilde{\mathbf{W}}^T$. Such objects will as a result of the 2PLS compression be removed from $\hat{\mathbf{X}}$ and totally captured in the new residual $\tilde{\mathbf{E}}$. Many objects may in fact have small projections on $\widetilde{\mathbf{w}}_2$ also if they have large projections to either $\hat{\mathbf{w}}_2$ or $\hat{\mathbf{w}}_3$, and thus more of $\mathbf{X}$ will be captured in $\tilde{\mathbf{E}}$ than in $\hat{\mathbf{E}}$. Note, however, that the estimator $\hat{\mathbf{b}}$ will be exactly the same.

3 Latent variable profile estimation

3.1 Introduction

An argumentation for keeping the first loading vector $\hat{\mathbf{w}}_1$ in the two 2PLS factorizations (9) and (11) may be based on profile estimation properties. For a spectral measurement on a mixture of pure constituents, $\hat{\mathbf{w}}_1$ will be a scaled LS estimate of the spectral profile of the constituent used as response variable. The second loading vector $\tilde{\mathbf{p}}_2^{\mathrm{W}}$ in the factorization (11) will furthermore be a scaled estimate of a weighted sum of the spectral profiles of all the interferants.

3.2 Latent variable modeling

Assume centered data generated according to the LV model

$$y_k = \mathbf{C}_1\mathbf{z}_k + f_k \tag{12}$$

$$\mathbf{x}_k = \mathbf{C}_2\mathbf{z}_k + \mathbf{e}_k, \tag{13}$$

where $\mathbf{z}_k = \left[\begin{array}{cccc} z_1 & z_2 & \dots & z_A \end{array}\right]^T \in \mathbb{R}^{A\times 1}$ is a random vector of latent variables, i.e. the expectation $E\mathbf{z}_j\mathbf{z}_k^T = \mathbf{0}$ for all $j \neq k$, and where y_k is a scalar response variable, while $\mathbf{x}_k \in \mathbb{R}^{p\times 1}$ is a vector of regressor variables. $\mathbf{C}_1 \in \mathbb{R}^{1\times A}$ and $\mathbf{C}_2 \in \mathbb{R}^{p\times A}$ are time-invariant matrices, while f_k and $\mathbf{e}_k$ are independent and random noise of appropriate dimensions.

Also assume independent latent variables in $\mathbf{z}_k$, i.e. a diagonal expectation $E\mathbf{z}_k\mathbf{z}_k^T$. Without loss of generality we may then assume an LV representation such that $\mathbf{C}_1 = \left[\begin{array}{cc} 1 & \mathbf{0} \end{array}\right]$, i.e. we assume that the response variable is a specific latent variable plus some random noise. Collection of data from N observations in a vector $\mathbf{y} \in \mathbb{R}^{N\times 1}$ and a matrix $\mathbf{X} \in \mathbb{R}^{N\times p}$ thus gives

$$\mathbf{y} = \mathbf{Z}\mathbf{C}_1^T + \mathbf{f} = \left[\begin{array}{cc} \mathbf{z}_1 & \mathbf{Z}_2 \end{array}\right]\left[\begin{array}{c} 1 \\ \mathbf{0} \end{array}\right] + \mathbf{f} = \mathbf{z}_1 + \mathbf{f} \tag{14}$$

$$\mathbf{X} = \mathbf{Z}\mathbf{C}_2^T + \mathbf{E} = \mathbf{z_1}\mathbf{C}_{21}^T + \mathbf{Z}_2\mathbf{C}_{22}^T + \mathbf{E} = \mathbf{y}\mathbf{C}_{21}^T - \mathbf{f}\mathbf{C}_{21}^T + \mathbf{Z}_2\mathbf{C}_{22}^T + \mathbf{E}, \tag{15}$$

where it is a part of the assumptions that $A << \min(N, p)$. The columns of $\mathbf{C}_2$ may typically be scaled versions of pure constituent spectral profiles. It also follows from the assumptions that the columns of $\mathbf{Z}_2$ are orthogonal to $\mathbf{y}$, i.e. all LV vectors except $\mathbf{z}_1$ are orthogonal to $\mathbf{y}$. Finally note that the model (14,15) can be transformed into the model (1,2) by a similarity transformation.

3.3 Profile estimation

3.3.1 Estimation of primary spectral profile

The column $\mathbf{C}_{21}$ of $\mathbf{C}_2$ that is directly related to the response $\mathbf{y}$ can be found from Eq. (15) using LS regression according to

$$\hat{\mathbf{C}}_{21} = \mathbf{X}^T\mathbf{y}\left(\mathbf{y}^T\mathbf{y}\right)^{-1}. \tag{16}$$

Under the assumptions given the underlying LS problem is well-posed. It is also a central part of the PLSR algorithms that the first loading weight vector with a scalar response variable is found according to Eq. (6). From Eqs. (6) and (16) thus follows that $\hat{\mathbf{C}}_{21}=\sqrt{\mathbf{y}^T\mathbf{X}\mathbf{X}^T\mathbf{y}}\left(\mathbf{y}^T\mathbf{y}\right)^{-1}\hat{\mathbf{w}}_1$. With the representation used in Eq. (14), i.e. $\mathbf{y} = \mathbf{z}_1 + \mathbf{f}$, the first loading weight vector $\hat{\mathbf{w}}_1$ thus gives a scaled LS estimate of $\mathbf{C}_{21}$. This is an argument for keeping $\hat{\mathbf{w}}_1$ also after the 2PLS compression.

3.3.2 Isolation of y-orthogonal components

From Eq. (15) follows

$$\mathbf{X} - \mathbf{y}\mathbf{C}_{21}^T = \mathbf{Z}_2\mathbf{C}_{22}^T - \mathbf{f}\mathbf{C}_{21}^T + \mathbf{E}. \tag{17}$$

Using $\hat{\mathbf{C}}_{21}$ from Eq. (16) we may compute $\mathbf{X} - \mathbf{y}\hat{\mathbf{C}}_{21}^T$, and singular value decomposition (SVD) of the result gives

$$\mathbf{X} - \mathbf{y}\hat{\mathbf{C}}_{21}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T = \begin{bmatrix} \mathbf{U}_2 & \mathbf{U}_\mathbf{E} \end{bmatrix} \begin{bmatrix} \mathbf{S}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_\mathbf{E} \end{bmatrix} \begin{bmatrix} \mathbf{V}_2^T \\ \mathbf{V}_\mathbf{E}^T \end{bmatrix} = \mathbf{U}_2\mathbf{S}_2\mathbf{V}_2^T + \mathbf{E}. \tag{18}$$

Choosing $\hat{\mathbf{Z}}_2 = \mathbf{U}_2$ we will thus find the confounded, scaled and noise corrupted profiles of the $\mathbf{y}$-orthogonal LVs in $\hat{\mathbf{C}}_{22} = \mathbf{V}_2\mathbf{S}_2$. Note that the scaled and sign indeterminate profile estimate of a single unknown LV will be found directly from this.

3.3.3 2PLS result

In the noise free case it follows from Eq. (15) and use of the orthogonal scores factorization (11) that

$$\mathbf{X} = \mathbf{y}\mathbf{C}_{21}^T + \mathbf{Z}_2\mathbf{C}_{22}^T = \tilde{\mathbf{t}}_1^\mathrm{W}\hat{\mathbf{w}}_1^T + \tilde{\mathbf{t}}_2^\mathrm{W}\left(\tilde{\mathbf{p}}_2^\mathrm{W}\right)^T, \tag{19}$$

where $\tilde{\mathbf{t}}_1^\mathrm{W}\hat{\mathbf{w}}_1^T = \mathbf{y}\mathbf{C}_{21}^T$, i.e. the $\tilde{\mathbf{p}}_2^\mathrm{W}$ vector is a weighted and noise corrupted sum of all columns in $\mathbf{C}_2$ in the model (12,13) except $\mathbf{C}_{21}$,

$$\tilde{\mathbf{p}}_2^\mathrm{W} \approx \mathbf{C}_{22}\mathbf{Z}_2^T\tilde{\mathbf{t}}_2^\mathrm{W}\left(\left(\tilde{\mathbf{t}}_2^\mathrm{W}\right)^T\tilde{\mathbf{t}}_2^\mathrm{W}\right)^{-1}. \tag{20}$$

Although some structured $\mathbf{X}$-variation may be captured in $\tilde{\mathbf{E}}$, Eq. (20) may still, together with Eq. (18), be an aid for the interpretation with regard to unknown LVs (see simulation example in Section 4).

3.3.4 Consequences of data centering and standardization

Centering of the data, i.e. using $\mathbf{X} \leftarrow \mathbf{X} - \bar{\mathbf{X}}$ and $\mathbf{y} \leftarrow \bar{\mathbf{y}}$, where $\bar{\mathbf{X}}$ and $\bar{\mathbf{y}}$ are column mean values, has no effect on the $\hat{\mathbf{C}}_2$ estimates according to Eqs. (16) and (18). However, standardization of the columns of $\mathbf{X}$ and $\mathbf{y}$ to unit variance does affect $\hat{\mathbf{C}}_2$, and must thus be properly accounted for.

4 Simulation example

The practical case behind the following simulation example could be a spectroscopic measurement of a solution with three different chemical constituents. A typical simulation result is shown in Fig. 2. Note the overlapping peaks and considerable **X**-noise.

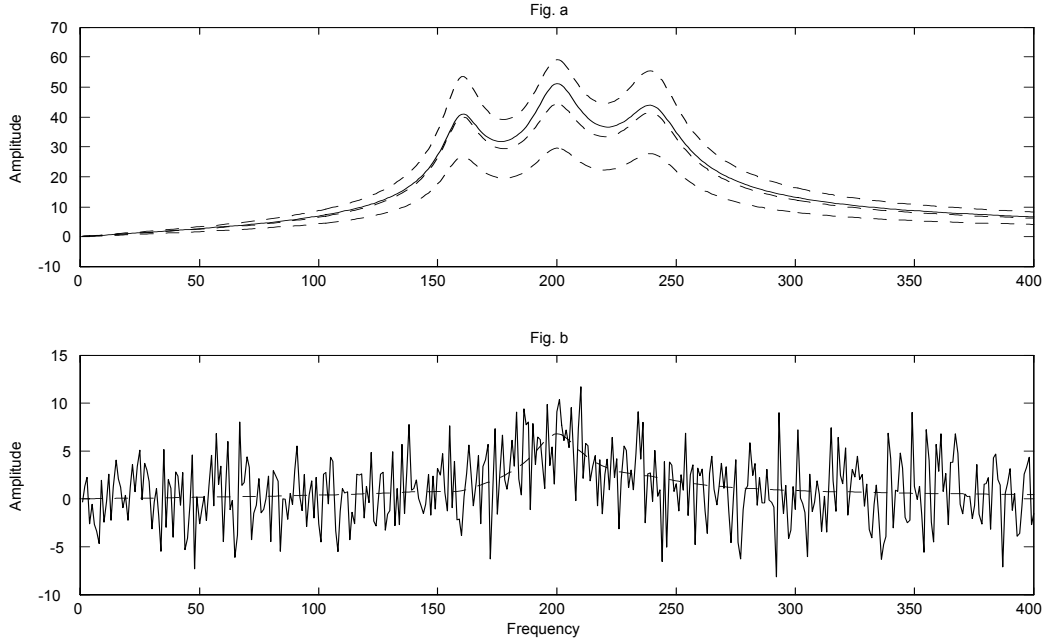

Figure 2. Mean spectrum and standard deviations (Fig. a - dashed lines) plus a typical realization of a noise free original spectrum (Fig. a - solid line), and a corresponding centered and noise corrupted spectrum (Fig. b) of a mixture of three chemical constituents. The centered noise free spectrum is shown by dashed line in Fig. b.

The simulations are based on assumed discrete frequency spectra in the range $0 < f \leq 400$ frequency units, constituting $\mathbf{C}_2$ in the LV model (12,13) (see [6] for details). It is assumed that the variations in the concentrations of the three constituents, $z_{1,k}$, $z_{2,k}$ and $z_{3,k}$, are independent and randomly generated zero mean numbers with normal distributions and variances $Ez_{1,k}^2 = Ez_{2,k}^2 = Ez_{3,k}^2 = 1$. The noise terms $e_k(f)$ are independent and randomly generated zero mean numbers with normal distribution and equal variances $Ee_k^2(f) = 10$. The response variable y_k is assumed to be the concentration z_2 of constituent 2, with a measurement error variance of $Ef_k^2 = 0.0001$. The optimal number of original PLS components using mean centered data, $N = 200$ modeling samples and $N_{val.} = 200$ independent validation samples were found to be $A = 3$, typically resulting in 91% explained y-variance. The PLSR and 2PLSR estimators (5) and (8) were exactly the same (using long Matlab format).

Examples of the resulting normalized estimates of the known constituent spectrum, and of the confounded unknown interferants spectra are shown in Fig. 3. Note, however, that parts of the interferants spectra may be captured in $\tilde{\mathbf{E}}$.

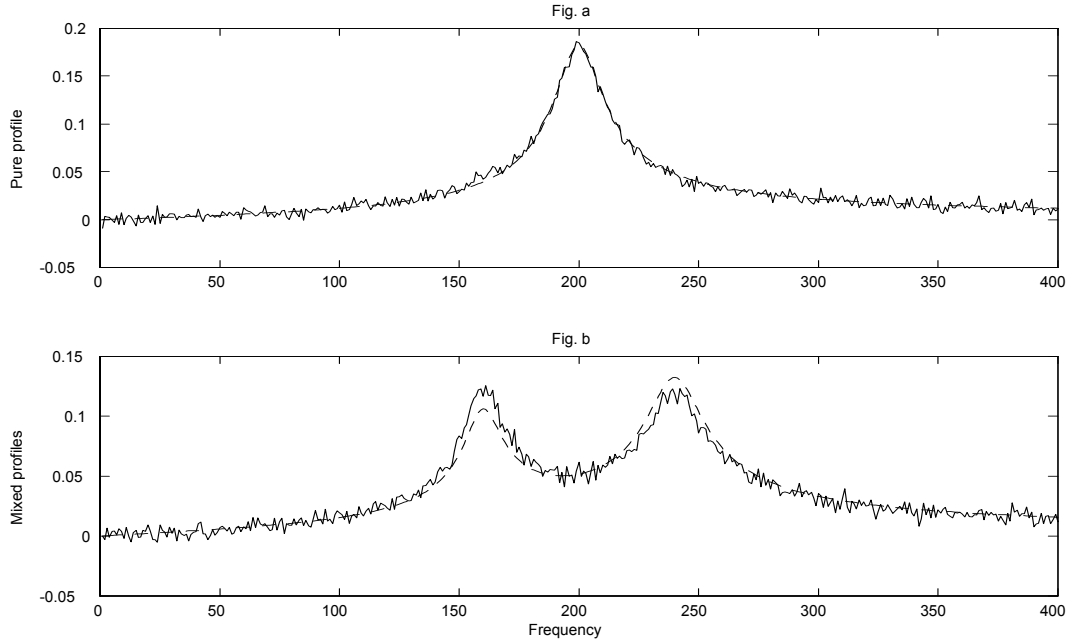

Figure 3. Normalized estimates of known constituent spectrum (Fig. a) and of confounded unknown interferants spectra (Fig. b). Solid lines show $\hat{\mathbf{w}}_1$ and $\tilde{\mathbf{p}}_2^{\mathrm{W}}$ in the orthogonal scores 2PLS factorization (11), while the dashed lines are given by Eqs. (16) and (20) after normalization.

The corresponding score plots for the ordinary orthogonal scores PLS model (3) are shown in Fig. 4a and 4b, while the score plot for the orthogonal scores 2PLS model (11) is shown in Fig. 4c. Fig. 4d shows $\hat{y}$ vs. y for the modeling set. Note that the specific o-marked sample might be questioned as an outlier (Fig. 4a), although Fig. 4d shows that the corresponding prediction is well in line. Using the two-component 2PLS model it can be seen directly in the score plot (Fig. 4c) that this sample gives a very small $\hat{y}$ value, although it would still be an outlier if the corresponding value of y was far from zero.

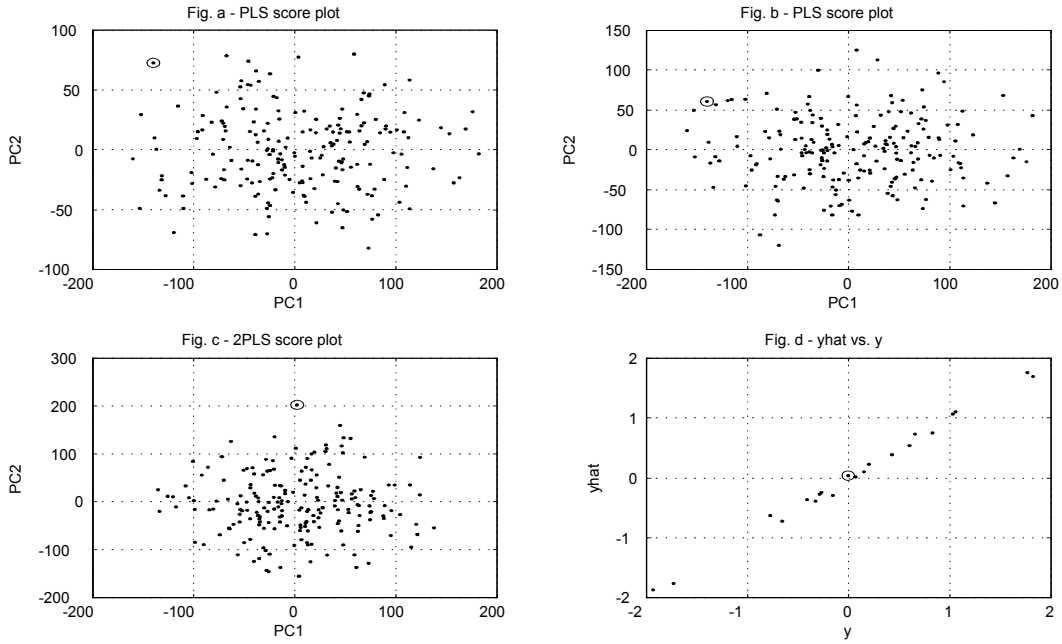

Figure 4. Score plots for ordinary orthogonal scores PLS model (Fig. a and b), and for orthogonal scores 2PLS model (Fig. c). Fig. d shows $\hat{y}$ vs. y (a limited number of samples only, for clarity). Sample 178 is o-marked in all plots.

Fig. 5 shows corresponding simulation results with two extra interferants. Note that the theoretical mixed spectrum (20) is using a very low weight for the last interferant ($f_5 = 280$), and a low weight also for the first interferant ($f_1 = 120$), as reflected in the mixed spectra results shown. However, these interferants are clearly visible in the residual spectra, determined as the three first columns in the matrix $\mathbf{V}$ in the SVD (i.e. principal component analysis)

$$\tilde{\mathbf{E}} = \mathbf{X} - \tilde{\mathbf{T}}_{\mathrm{W}} \tilde{\mathbf{P}}_{\mathrm{W}}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T. \tag{21}$$

The other columns in $\mathbf{V}$ showed no structured variation. Note that also these residual spectra are normalized, i.e. the actual amplitudes cannot be judged from the plot.

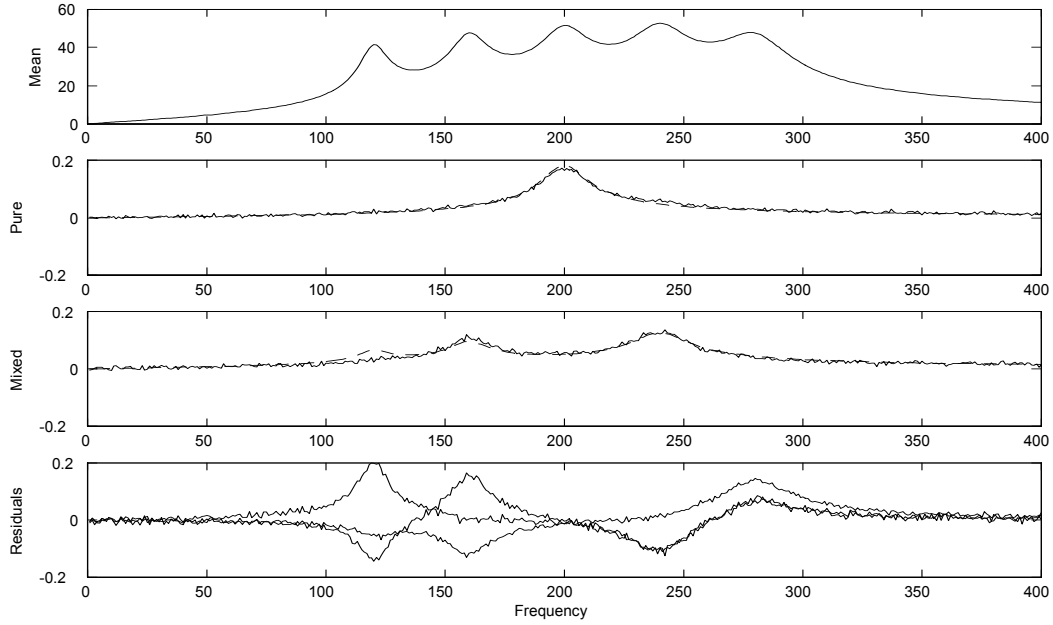

Figure 5. Results from simulations with two added interferants. Compare with the mean noise free spectrum in Fig. 2, the pure profile in Fig. 3a and the mixed profiles in Fig. 3b. The last plot is determined by a SVD/PCA of the residuals from the 2PLS model.

5 Industrial data example

The data in this example is with permission from the company borrowed from the function *plsdemo* in the Eigenvector PLS Toolbox 2.0.1b [7]. Each sample consists of 20 temperature measurements and a level measurement from a Liquid-Fed Ceramic Melter. As is typical for many industrial process cases, the $\mathbf{X}$ and $\mathbf{y}$ data are here time series, i.e. they have dependent samples. As in the toolbox the data are used to develop a model that relates temperatures $\mathbf{x}^T$ in a molten glass tank to the tank level y. The data is given in two independent time series, with 300 samples in Block 1 and 200 samples in Block 2. The main point in the present example is a comparison between an ordinary PLS model and the 2PLS model developed in Section 2, related to outlier detection and fault detection.

Fig. 6 shows results using all 300 samples in the Block 1 data to find first an ordinary orthogonal scores PLS model with mean centered data and $A = 3$ components, and based on that an orthogonal scores 2PLS model giving exactly the same estimator $\hat{\mathbf{b}}$. The first score plot (PC2 vs. PC1) from the PLS model (Fig. 6a) shows a rather obvious o-marked outlier (sample 73). However, this outlier has a very insignificant effect on the predictions when validated against the Block 2 data (58.1 % explained y-variance with sample 73 included, and 58.2 % without), indicating some largely $\mathbf{y}$-orthogonal features. The score plot in Fig. 6c is based on the orthogonal scores 2PLS model (11), and the o-marked sample now looks much more normal, indicating that the $\mathbf{y}$-orthogonal feature has been filtered out and captured in the residuals $\tilde{\mathbf{E}}$. The ◊-marked sample now looks more like an outlier, although not in an extreme way. Fig. 6b and d show $\hat{y}$ results using the the estimator $\hat{\mathbf{b}}$ on samples 51 to 110 in the modeling set. The two o- and ◊-marked samples here appear quite normal, while a third □-marked sample now appears the most abnormal. This last sample appears quite normal in the two score plots, and the large prediction error is most likely caused by an error in the y-measurement.

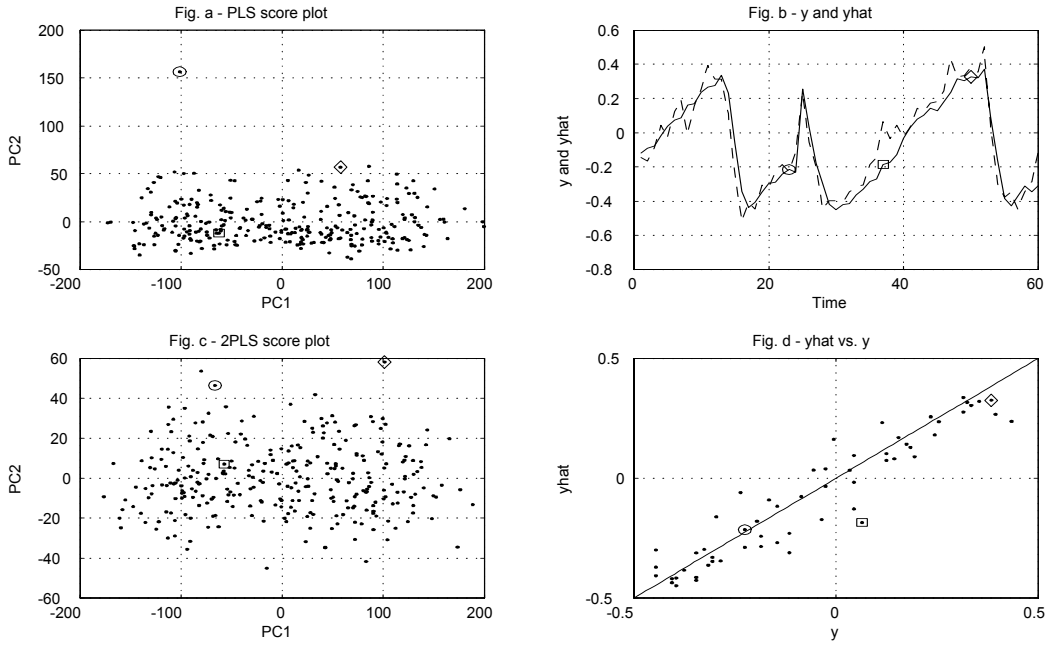

Figure 6. Results when using samples 1-300 for PLS and 2PLS modeling and outlier detection. The 2PLS model largely overlooks the apparent o-marked outlier resulting from ordinary PLS modeling, which may be a good feature since the predictor is not affected. The solid line in Fig. 6b is $\hat{y}$. See text for other details.

In the next step samples 131 to 300 were used for modeling, while samples 51-110 were used for testing fault detection properties. Also here orthogonal scores PLS and 2PLS models were used. Fig. 7a shows the first PLS score plot (PC2 vs. PC1) for the test set. A solid line is used up to sample 73 (while the dashed line indicates the process movements ahead of sample 73), and this would be the picture if the score plot was used for process monitoring. The large deviation of sample 73 would cause some concern, although such a one-sample departure would be overlooked by a somewhat conservative fault detection system. The corresponding deviation is much smaller in the 2PLS score plot in Fig. 7c. Here, the ◊-marked sample would cause more concern, although the deviation is considerably less than for sample 73 in Fig. 7a. For the ordinary PLS the ◊-marked deviation is most pronounced in the second score plot (PC3 vs. PC1) in Fig. 7b. However, since vertical movements in the 2PLS score plot are **y**-orthogonal, we can conclude that both the o- and the ◊-marked one-sample deviations have very little to do with variations in y. The $\hat{y}$ vs. y plot is very much the same as in Fig. 6d, and is therefore not included. Instead a PCP plot according to [3] is included in Fig. 7d, scaled to the same score variances as for the 2PLS plot. The o- and the ◊-marked one-sample deviations are also here **y**-orthogonal, but since the second score vector is found from the first principal component of the residual after removal of the first (and only) PCP component from **X**, the deviation for sample 73 is just as pronounced as in Fig. 7a.

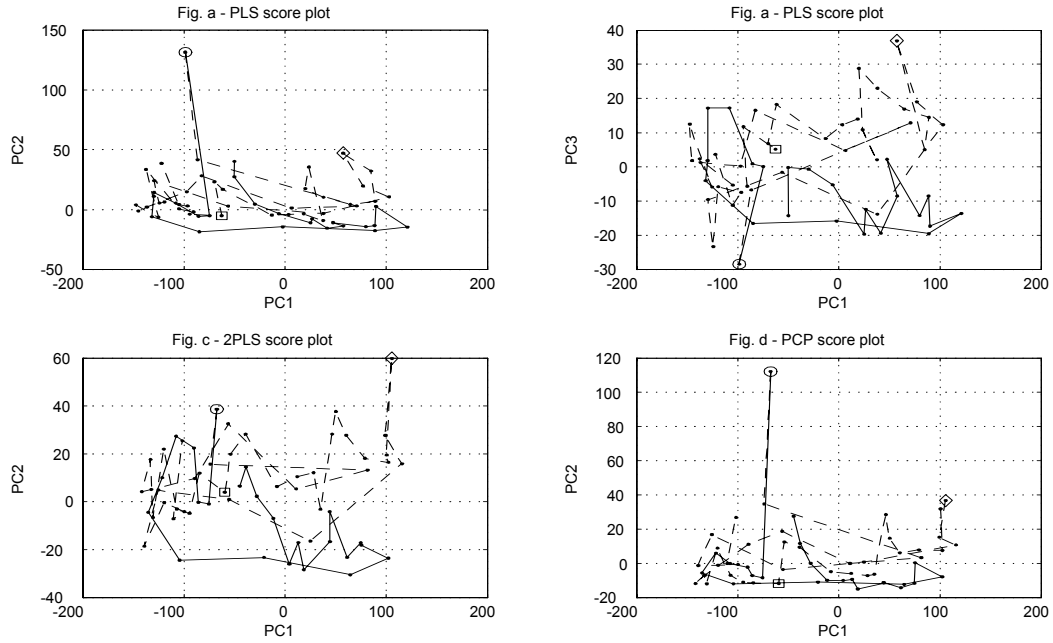

Figure 7. Results when using samples 131-300 for PLS and 2PLS modeling, and samples 51-110 for testing and fault detection. The 2PLS score plot shows that the o- and ◊-marked one-sample departures are **y**-orthogonal. The maximal departure is also much smaller in the 2PLS score plot than in the PLS score plots. Fig. d shows a PCP plot for comparison. See text for details.

The example shows that compression from an ordinary PLS model into a two-component 2PLS model may filter out **y**-orthogonal variation in $\mathbf{X}$ in such a way that score plot based outlier detection (Fig. 6) and fault detection (Fig. 7) are made more response relevant. The possibility to distinguish between **y**-related and **y**-orthogonal movements in the 2PLS score plot gives additional interpretational power.

6 Conclusions

Assuming a scalar response variable, all PLS factorizations with any number of components can by use of a simple algorithm be compressed into two components only. Different 2PLS transformations for use in different applications are possible. The advantage with a complete two-component model is easier interpretation in spectra estimation, outlier detection, fault detection etc., as exemplified in simulations and an industrial data case. A process monitoring application using score-loading correspondence will be reported separately.

A Proof of Theorem 1

Lemma 1 Given the factorization (4) and $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}$ we have

$$\mathbf{T}_{\mathrm{M}}^T(\mathbf{y}-\hat{\mathbf{y}}) = \hat{\mathbf{W}}^T\mathbf{X}^T(\mathbf{y}-\hat{\mathbf{y}}) = \hat{\mathbf{W}}^T\mathbf{X}^T\left(\mathbf{y}-\hat{\mathbf{X}}\hat{\mathbf{W}}\left(\hat{\mathbf{W}}^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{W}}\right)^{-1}\hat{\mathbf{W}}^T\mathbf{X}^T\mathbf{y}\right) = \mathbf{0}, \qquad (22)$$

i.e. $\mathbf{y}-\hat{\mathbf{y}}$ is orthogonal to the score vectors in $\mathbf{T}_{\mathrm{M}} = \mathbf{X}\hat{\mathbf{W}}$ (and thus also to the score vectors in $\mathbf{T}_{\mathrm{W}}$).

Lemma 2 By insertion of $\hat{\mathbf{b}} = \hat{\mathbf{W}}\left(\hat{\mathbf{W}}^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{W}}\right)^{-1}\hat{\mathbf{W}}^T\mathbf{X}^T\mathbf{y}$ and

$$\hat{\mathbf{b}}_1 = \hat{\mathbf{w}}_1\left(\hat{\mathbf{w}}_1^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{w}}_1\right)^{-1}\hat{\mathbf{w}}_1^T\mathbf{X}^T\mathbf{y} \tag{23}$$

(the estimator when one component only is used) it is straightforward to show that

$$\hat{\mathbf{b}}\left(\hat{\mathbf{b}}^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}}\right)^{-1}\hat{\mathbf{b}}^T\mathbf{X}^T\mathbf{y} = \hat{\mathbf{b}}, \tag{24}$$

$$\hat{\mathbf{b}}_1\left(\hat{\mathbf{b}}_1^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}}_1\right)^{-1}\hat{\mathbf{b}}_1^T\mathbf{X}^T\mathbf{y} = \hat{\mathbf{b}}_1. \tag{25}$$

From this also follows $\hat{\mathbf{b}}^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}} = \hat{\mathbf{b}}^T\mathbf{X}^T\mathbf{y}$ and $\hat{\mathbf{b}}_1^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}}_1 = \hat{\mathbf{b}}_1^T\mathbf{X}^T\mathbf{y}$.

Proof of Theorem 1 Just as the factorization (4) gives the estimator (5), the factorization (9) gives the estimator (8) (independent of the number of components, see [6]). It remains to show that the two estimators are identical. As a first step we perform a similarity transformation of (9) (see subsection 2.3 for principle) into

$$\mathbf{X} = \tilde{\mathbf{t}}_1\hat{\mathbf{w}}_1^T + \tilde{\mathbf{t}}_2\tilde{\mathbf{p}}_2^T + \tilde{\mathbf{E}}, \tag{26}$$

where $\tilde{\mathbf{p}}_2$ is a unity vector in the plane defined by $\hat{\mathbf{w}}_1$, $\hat{\mathbf{b}}$ and $\widetilde{\mathbf{w}}_2$ (see Fig. 1). More precisely we specify

$$\tilde{\mathbf{p}}_2 = \frac{\hat{\mathbf{b}} - \hat{\mathbf{b}}_1}{\left\|\hat{\mathbf{b}} - \hat{\mathbf{b}}_1\right\|} = \frac{\hat{\mathbf{b}}_2}{\left\|\hat{\mathbf{b}}_2\right\|}. \tag{27}$$

The loading matrix is now $\tilde{\mathbf{P}} = \begin{bmatrix} \hat{\mathbf{w}}_1 & \tilde{\mathbf{p}}_2 \end{bmatrix}$, and the estimator (8) thus becomes

$$\tilde{\mathbf{b}} = \tilde{\mathbf{P}}\left(\tilde{\mathbf{P}}^T\mathbf{X}^T\mathbf{X}\tilde{\mathbf{P}}\right)^{-1}\tilde{\mathbf{P}}^T\mathbf{X}\mathbf{y}. \tag{28}$$

It still remains to show that $\tilde{\mathbf{b}}$ is identical to $\hat{\mathbf{b}}$ according to Eq. (6).

The detailed development of $\tilde{\mathbf{b}}$ is facilitated by the fact that $\hat{\mathbf{w}}_1^T\mathbf{X}^T\mathbf{X}\tilde{\mathbf{p}}_2 = 0$. This is so since after introduction of $\hat{\mathbf{y}}_1 = \mathbf{X}\hat{\mathbf{b}}_1$ we have

$$\begin{aligned}\hat{\mathbf{w}}_1^T\mathbf{X}^T\mathbf{X}\tilde{\mathbf{p}}_2\|\tilde{\mathbf{b}}_2\| &= \hat{\mathbf{w}}_1^T\mathbf{X}^T\mathbf{X}\tilde{\mathbf{b}}_2 = \hat{\mathbf{w}}_1^T\mathbf{X}^T\left(\mathbf{X}\hat{\mathbf{b}} - \mathbf{X}\hat{\mathbf{b}}_1\right) = \hat{\mathbf{w}}_1^T\mathbf{X}^T\left(\hat{\mathbf{y}} - \hat{\mathbf{y}}_1\right) \\ &= -\mathbf{t}_1^M\left[(\mathbf{y} - \hat{\mathbf{y}}) - (\mathbf{y} - \hat{\mathbf{y}}_1)\right] = 0,\end{aligned} \tag{29}$$

where the last equality follows from Lemma 1 because $\mathbf{y} - \hat{\mathbf{y}}_1$ and $\mathbf{y} - \hat{\mathbf{y}}$ are both orthogonal to $\tilde{\mathbf{t}}_1^{\mathrm{M}} = \mathbf{X}\hat{\mathbf{w}}_1$. Since $\hat{\mathbf{w}}_1$ and $\tilde{\mathbf{p}}_2$ have the same directions as $\hat{\mathbf{b}}_1$ and $\hat{\mathbf{b}}_2$, this also gives $\hat{\mathbf{b}}_1^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}}_2 = 0$, and thus

$$\hat{\mathbf{b}}^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}}_1 = \left(\hat{\mathbf{b}}_1 + \hat{\mathbf{b}}_2\right)^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}}_1 = \hat{\mathbf{b}}_1^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}}_1 + \hat{\mathbf{b}}_2^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}}_1 = \hat{\mathbf{b}}_1^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}}_1. \tag{30}$$

From Eqs. (24), (25), (27), (28) and (30) and use of Lemma 2 thus follows

$$\begin{aligned}\tilde{\mathbf{b}} &= \hat{\mathbf{w}}_1\left(\hat{\mathbf{w}}_1^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{w}}_1\right)^{-1}\hat{\mathbf{w}}_1^T\mathbf{X}^T\mathbf{y} + \tilde{\mathbf{p}}_2\left(\tilde{\mathbf{p}}_2^T\mathbf{X}^T\mathbf{X}\tilde{\mathbf{p}}_2\right)^{-1}\tilde{\mathbf{p}}_2^T\mathbf{X}^T\mathbf{y} \\ &= \hat{\mathbf{b}}_1 + \hat{\mathbf{b}}_2\left(\hat{\mathbf{b}}_2^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}}_2\right)^{-1}\hat{\mathbf{b}}_2^T\mathbf{X}^T\mathbf{y} \\ &= \hat{\mathbf{b}}_1 + \hat{\mathbf{b}}_2\left(\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_1\right)^T\mathbf{X}^T\mathbf{X}\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_1\right)\right)^{-1}\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_1\right)^T\mathbf{X}^T\mathbf{y} \\ &= \hat{\mathbf{b}}_1 + \hat{\mathbf{b}}_2\left(\hat{\mathbf{b}}^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}} - 2\hat{\mathbf{b}}^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}}_1 + \hat{\mathbf{b}}_1^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}}_1\right)^{-1}\left(\hat{\mathbf{b}}^T\mathbf{X}^T\mathbf{y} - \hat{\mathbf{b}}_1^T\mathbf{X}^T\mathbf{y}\right) \\ &= \hat{\mathbf{b}}_1 + \hat{\mathbf{b}}_2\left(\hat{\mathbf{b}}^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}} - \hat{\mathbf{b}}_1^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}}_1\right)^{-1}\left(\hat{\mathbf{b}}^T\mathbf{X}^T\mathbf{y} - \hat{\mathbf{b}}_1^T\mathbf{X}^T\mathbf{y}\right) = \hat{\mathbf{b}}_1 + \hat{\mathbf{b}}_2 = \hat{\mathbf{b}}.\end{aligned} \tag{31}$$

The score expressions $\hat{\mathbf{t}}_1 = \mathbf{X}\hat{\mathbf{w}}_1$ and $\tilde{\mathbf{t}}_2^{\mathrm{M}} = \mathbf{X}\widetilde{\mathbf{w}}_2$ follow from Eq. (26) since $\hat{\mathbf{w}}_1$ and $\widetilde{\mathbf{w}}_2$ are orthogonal, with $\hat{\mathbf{w}}_1^T\hat{\mathbf{w}}_1 = 1$ and $\widetilde{\mathbf{w}}_2^T\widetilde{\mathbf{w}}_2 = 1$. It remains to prove that $\mathbf{y}^T\tilde{\mathbf{t}}_2^{\mathrm{M}} = 0$ and $\hat{\mathbf{y}}^T\tilde{\mathbf{t}}_2^{\mathrm{M}} = 0$. From Eq. (5) follows

$$\mathbf{y}^T\tilde{\mathbf{t}}_2^{\mathrm{M}} = \mathbf{y}^T\mathbf{X}\widetilde{\mathbf{w}}_2 = c_1^{-1}\hat{\mathbf{w}}_1^T\widetilde{\mathbf{w}}_2 = 0, \tag{32}$$

which with use of Eq. (8) further gives

$$\begin{aligned}
\hat{\mathbf{y}}^T\tilde{\mathbf{t}}_2^{\mathrm{M}} &= \hat{\mathbf{b}}^T\mathbf{X}^T\tilde{\mathbf{t}}_2^{\mathrm{M}} = \mathbf{y}^T\mathbf{X}\tilde{\mathbf{W}}\left(\tilde{\mathbf{W}}^T\mathbf{X}^T\mathbf{X}\tilde{\mathbf{W}}\right)^{-1}\tilde{\mathbf{W}}^T\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{w}}_2 \\
&= \mathbf{y}^T\begin{bmatrix}\hat{\mathbf{t}}_1 & \tilde{\mathbf{t}}_2^{\mathrm{M}}\end{bmatrix}\begin{bmatrix}\hat{\mathbf{w}}_1^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{w}}_1 & \hat{\mathbf{w}}_1^T\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{w}}_2 \\ \hat{\mathbf{w}}_1^T\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{w}}_2 & \widetilde{\mathbf{w}}_2^T\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{w}}_2\end{bmatrix}^{-1}\begin{bmatrix}\hat{\mathbf{w}}_1^T\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{w}}_2 \\ \widetilde{\mathbf{w}}_2^T\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{w}}_2\end{bmatrix} \\
&= \frac{1}{\hat{\mathbf{w}}_1^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{w}}_1\widetilde{\mathbf{w}}_2^T\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{w}}_2 - \left(\hat{\mathbf{w}}_1^T\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{w}}_2\right)^2}\begin{bmatrix}\mathbf{y}^T\hat{\mathbf{t}}_1 & 0\end{bmatrix} \\
&\quad\times\begin{bmatrix}\widetilde{\mathbf{w}}_2^T\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{w}}_2 & -\hat{\mathbf{w}}_1^T\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{w}}_2 \\ -\hat{\mathbf{w}}_1^T\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{w}}_2 & \hat{\mathbf{w}}_1^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{w}}_1\end{bmatrix}\begin{bmatrix}\hat{\mathbf{w}}_1^T\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{w}}_2 \\ \widetilde{\mathbf{w}}_2^T\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{w}}_2\end{bmatrix} \\
&= \text{const.}\ \times\ \left[\left(\widetilde{\mathbf{w}}_2^T\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{w}}_2\right)\left(\hat{\mathbf{w}}_1^T\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{w}}_2\right) - \left(\hat{\mathbf{w}}_1^T\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{w}}_2\right)\left(\widetilde{\mathbf{w}}_2^T\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{w}}_2\right)\right] = 0.
\end{aligned} \tag{33}$$

B Details of orthogonal scores 2PLS factorization

The transformation

$$\begin{bmatrix}\tilde{\mathbf{t}}_1^{\mathrm{W}} & \tilde{\mathbf{t}}_2^{\mathrm{W}}\end{bmatrix}\begin{bmatrix}\hat{\mathbf{w}}_1^T \\ \left(\tilde{\mathbf{p}}_2^{\mathrm{W}}\right)^T\end{bmatrix} = \begin{bmatrix}\hat{\mathbf{t}}_1 & \tilde{\mathbf{t}}_2^{\mathrm{M}}\end{bmatrix}\begin{bmatrix}1 & 0 \\ d & f\end{bmatrix}\begin{bmatrix}1 & 0 \\ -d/f & 1/f\end{bmatrix}\begin{bmatrix}\hat{\mathbf{w}}_1^T \\ \widetilde{\mathbf{w}}_2^T\end{bmatrix} \tag{34}$$

gives $\tilde{\mathbf{t}}_1^{\mathrm{W}} = \hat{\mathbf{t}}_1 + d\tilde{\mathbf{t}}_2^{\mathrm{M}}$ and $\tilde{\mathbf{t}}_2^{\mathrm{W}} = f\tilde{\mathbf{t}}_2^{\mathrm{M}}$. In order to obtain $\left(\tilde{\mathbf{t}}_1^{\mathrm{W}}\right)^T\tilde{\mathbf{t}}_2^{\mathrm{W}} = \left(\hat{\mathbf{t}}_1 + d\tilde{\mathbf{t}}_2^{\mathrm{M}}\right)^T f\tilde{\mathbf{t}}_2^{\mathrm{M}} = 0$, we must chose $d = -\hat{\mathbf{t}}_1^T\tilde{\mathbf{t}}_2^{\mathrm{M}}/\left(\tilde{\mathbf{t}}_2^{\mathrm{M}}\right)^T\tilde{\mathbf{t}}_2^{\mathrm{M}}$. We further find $\tilde{\mathbf{p}}_2^{\mathrm{W}} = \left(-d\hat{\mathbf{w}}_1 + \widetilde{\mathbf{w}}_2\right)/f$, and in order to get $\left(\tilde{\mathbf{p}}_2^{\mathrm{W}}\right)^T\tilde{\mathbf{p}}_2^{\mathrm{W}} = 1$, we chose $f = \sqrt{1+d^2}$.

Finally, it remains to prove that $\tilde{\mathbf{t}}_1^{\mathrm{W}} = c\hat{\mathbf{y}}$ holds also for new objects. From the transformation above follows that

$$\tilde{t}_{1,new}^{\mathrm{W}} = \begin{bmatrix}\hat{t}_1 & \tilde{t}_2^{\mathrm{M}}\end{bmatrix}_{new}\begin{bmatrix}1 \\ d\end{bmatrix} = \mathbf{x}_{new}^T\tilde{\mathbf{W}}\begin{bmatrix}1 \\ d\end{bmatrix} = \mathbf{x}_{new}^T\tilde{\mathbf{W}}\begin{bmatrix}1 \\ -\hat{\mathbf{t}}_1^T\tilde{\mathbf{t}}_2^{\mathrm{M}}\left(\left(\tilde{\mathbf{t}}_2^{\mathrm{M}}\right)^T\tilde{\mathbf{t}}_2^{\mathrm{M}}\right)^{-1}\end{bmatrix}, \tag{35}$$

while the estimator (8) with use of the fact that $\left(\tilde{\mathbf{t}}_2^{\mathrm{M}}\right)^T\mathbf{y} = 0$ gives

$$\begin{aligned}
\hat{y}_{new} &= \mathbf{x}_{new}^T\tilde{\mathbf{W}}\left(\tilde{\mathbf{W}}^T\mathbf{X}^T\mathbf{X}\tilde{\mathbf{W}}\right)^{-1}\tilde{\mathbf{W}}^T\mathbf{X}^T\mathbf{y} = \mathbf{x}_{new}^T\tilde{\mathbf{W}}\begin{bmatrix}\hat{\mathbf{t}}_1^T\hat{\mathbf{t}}_1 & \hat{\mathbf{t}}_1^T\tilde{\mathbf{t}}_2^{\mathrm{M}} \\ \hat{\mathbf{t}}_1^T\tilde{\mathbf{t}}_2^{\mathrm{M}} & \left(\tilde{\mathbf{t}}_2^{\mathrm{M}}\right)^T\tilde{\mathbf{t}}_2^{\mathrm{M}}\end{bmatrix}^{-1}\begin{bmatrix}\hat{\mathbf{t}}_1^T\mathbf{y} \\ \left(\tilde{\mathbf{t}}_2^{\mathrm{M}}\right)^T\mathbf{y}\end{bmatrix} \\
&= \frac{1}{c}\mathbf{x}_{new}^T\tilde{\mathbf{W}}\begin{bmatrix}1 \\ -\hat{\mathbf{t}}_1^T\tilde{\mathbf{t}}_2^{\mathrm{M}}\left(\left(\tilde{\mathbf{t}}_2^{\mathrm{M}}\right)^T\tilde{\mathbf{t}}_2^{\mathrm{M}}\right)^{-1}\end{bmatrix} = \frac{1}{c}\tilde{t}_{1,new}^{\mathrm{W}}.
\end{aligned} \tag{36}$$

C Matlab code for 2PLS compression and transformation

```
%% Assume X, y and W known and find orthogonal loadings 2PLS factorization
a=inv(W'*X'*X*W)*W'*X'*y;
w1=W(:,1);
w2tilde=W(:,2:A)*a(2:A)/norm(W(:,2:A)*a(2:A));
t1=X*w1;
tm2tilde=X*w2tilde;
```

```
%% Transform to orthogonal scores 2PLS form
d=-t1'*tm2tilde/(tm2tilde'*tm2tilde);
f=sqrt(1+d^2);
p2tilde=(-d*w1+w2tilde)/f;
tw1tilde=t1+d*tm2tilde;
tw2tilde=f*tm2tilde;
```

References

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

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

[3] Langsrud Ø, Næs T. Optimised score plot by principal components of prediction, to appear in *Chemometrics Intell. Lab. Syst.* 2003.

[4] Helland IS. On the structure of partial least squares regression. *Communications in statistics* 1988; **17**: 581-607.

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

[6] Ergon R. PCR/PLSR optimization based on noise covariance estimation and Kalman filtering theory. *J. Chemometrics* 2002; **16**: 401-407.

[7] http://www.eigenvector.com