ELSEVIER

Chemometrics and Intelligent Laboratory Systems 56 (2001) 13–25

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

Direct orthogonal signal correction

Johan A. Westerhuis [a,*], Sijmen de Jong [b], Age K. Smilde [a]

[a] *Process Analysis and Chemometrics, Department of Chemical Engineering, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, Netherlands*

[b] *Unilever Research Vlaardingen, PO Box 114, 3130 AC Vlaardingen, Netherlands*

Received 4 July 2000; accepted 12 December 2000

Abstract

In the present paper, the concept of orthogonal signal correction (OSC) as a spectral preprocessing method is discussed and a number of OSC algorithms that have appeared are compared from a theoretical viewpoint. Since all of these algorithms had some problems concerning the orthogonality towards **Y**, non-optimal amount of variance removed from **X**, or a non-attainable solution, a new direct OSC algorithm (DOSC) is introduced. DOSC was originally developed as a direct method solely based on least squares steps that had none of the problems mentioned above. The first practical results with the new method, however, were not encouraging due to the complete orthogonality constraint. If this orthogonality constraint is loosened, the method improves considerably and simplifies the calibration model for the prediction of **Y**.

Keywords: Orthogonal signal correction; Near-infrared spectroscopy; PLS; Spectral preprocessing

1. Introduction

In spectroscopic calibrations where a partial least squares (PLS) or principal component regression (PCR) calibration model is used to predict a product quality such as a concentration or octane number, it is often encountered that the first component (or latent variable) accounts for a very high percentage of the variation in the spectral data **X** and only a low percentage of variation of the product quality **Y**. If more components are calculated the calibration model slowly improves. However, models with a large number of components are not desirable in terms of interpretability and robustness.

To deal with this problem, Wold et al. [1] introduced orthogonal signal correction (OSC). The goal of OSC is to remove one or more directions in **X**, orthogonal to **Y** that account for the largest variation in **X**. OSC is performed as a pre-processing step to improve the calibration model:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{1.1}$$

In this work, improving the calibration model is meant in a broad sense, such that the model is more parsimonious or that lower prediction errors of **Y** are obtained. The OSC method is almost always used together with a latent variable method such as PLS or PCR to build the calibration model.

* Corresponding author.
E-mail address: westerhuis@its.chem.uva.nl (J.A. Westerhuis).

0169-7439/01/$ - see front matter
PII: S0169-7439(01)00102-2

In Section 5.2 of the paper by Wold et al. [1], a remark about a simpler method is given. An optimal OSC would be to calculate principal components of the matrix **Z**, where **Z** is the **X**-matrix orthogonalized to **Y**. These components should describe the largest variation in **X** that is orthogonal to **Y**.

$$\mathbf{Z} = \mathbf{X} - \mathbf{Y}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{X} \quad (1.2)$$

or, in case **Y** is not of full column rank,

$$\mathbf{Z} = \mathbf{X} - \mathbf{Y}\mathbf{Y}^{+}\mathbf{X} \quad (1.3)$$

Here, $\mathbf{Y}^{+}$ is the Moore–Penrose inverse of **Y**. Such an approach will lead to a direct method in which, without iterations, orthogonal components are obtained as the principal components from **Z**. However, it was left to others to work out this approach because no solution was found for the fact that no **Y** values are available for future samples, and therefore, no orthogonalization can be performed. Also, there is no guarantee that **Z** lies in a subspace of **X**, and removing part of **Z**, e.g. its first PC, from **X** may introduce components outside the X-space into the corrected matrix.

Since the introduction of the OSC method by Wold et al., a number of different attempts to improve the OSC method have been presented in literature [2–4]. Furthermore, MATLAB code for an OSC algorithm has been published on the internet [5].

In the present paper, a direct orthogonal signal correction method (DOSC) is presented that calculates directions in **X** that are orthogonal to **Y** and account for the largest variance of **X**. These directions are obtained by only using least squares steps and they provide a theoretically exact solution to the problem set out by Wold.

In Section 2, first the different OSC methods introduced in the literature are discussed. Similarities and differences in these methods are addressed. Then the new DOSC method is introduced and its properties will be presented. In Section 4, two example data sets are used to show how DOSC compares to the other methods. This paper is not meant as a full comparison, but it shows some important properties of the methods.

2. Theory

In this section, the various OSC approaches presented in the literature are discussed. These are the approaches by Wold et al. [1], Sjöblom et al. [2], Wise and Gallagher [5], Andersson [3] and Fearn [4]. These names will be used for the corresponding approaches. For all cases, it is assumed that $\mathbf{X}(I \times J)$ contains the descriptive variables that are used to predict the response $\mathbf{Y}(I \times K)$. Throughout the paper, it is assumed that **X** and **Y** have been column-mean centered. The algorithms will only be discussed until the step where the OSC component is deflated from **X**, and only one OSC component has been calculated. For notational convenience, $\mathbf{P}_{\mathrm{D}}$ is defined as the orthogonal projector onto the column space of **D**, i.e. $\mathbf{P}_{\mathrm{D}} = \mathbf{D}\mathbf{D}^{+}$, and $\mathbf{A}_{\mathrm{D}}$ as the anti-projector with respect to D-space: $\mathbf{A}_{\mathrm{D}} = \mathbf{I} - \mathbf{P}_{\mathrm{D}} = \mathbf{I} - \mathbf{D}\mathbf{D}^{+}$. Furthermore, the notation $\mathrm{PC}_1(\mathbf{X})$ is used to denote the first principal component score vector of **X**.

2.1. Wold et al.

The approach presented by Wold et al. can be described in the following steps:

1. $\mathbf{t} = \mathrm{PC}_1(\mathbf{X})$
2. $\mathbf{t}^{*} = \mathbf{A}_{\mathrm{Y}}\mathbf{t}$
3. $\mathbf{t} = \mathbf{T}\mathbf{T}^{+}\mathbf{t}^{*}$

In step 3, a many-component PLS model between **X** and $\mathbf{t}^{*}$ is built. This means that $\mathbf{t}^{*}$ is projected on the scores **T** of the PLS model using the projection matrix $\mathbf{T}\mathbf{T}^{+}$ or $\mathbf{P}_{\mathrm{T}}$. This projection matrix, however, is only defined if $\mathbf{t}^{*}$ (which changes in every iteration) and the number of PLS components are defined. Repeat steps 2–3 until convergence of **t**, which is the score vector of the OSC component.

4. $\mathbf{p} = \mathbf{t}^{\mathrm{T}}\mathbf{X}/(\mathbf{t}^{\mathrm{T}}\mathbf{t})$
5. $\mathbf{X}^{\mathrm{Wold}} = \mathbf{X} - \mathbf{t}\mathbf{p}^{\mathrm{T}}$

If more OSC components are needed, steps 1–5 can be repeated again on $\mathbf{X}^{\mathrm{Wold}}$. For spectra of new

samples $\mathbf{x}_{new}$, the correction is performed by repeating the following steps for each OSC component i:

6. $t_{new} = \mathbf{x}_{new}^{T}\mathbf{b}_i$
7. $\mathbf{x}_{new}^{Wold} = \mathbf{x}_{new} - t_{new}\mathbf{p}_i$

where $\mathbf{b}_i$ is a vector of regression coefficients of the many-component PLS model in step 3 for the ith OSC component. In cycling through steps 2–3, an eigenvector of $\mathbf{P}_T\mathbf{A}_Y$ is found satisfying, $\mathbf{P}_T\mathbf{A}_Y\mathbf{t} = \mathbf{t}$, or $\mathbf{P}_T\mathbf{A}_Y\mathbf{t} = \lambda\mathbf{t}$ if a scaling of the PLS regression coefficient in step 3 is assumed.

For a better understanding of the approach by Wold et al., consider that in step 3, instead of many components, all PLS components would have been used, then $\mathbf{TT}^{+} = \mathbf{XX}^{+}$ or $\mathbf{P}_T = \mathbf{P}_X$. The solution of the OSC procedure then satisfies $\mathbf{P}_X\mathbf{A}_Y\mathbf{t} = \mathbf{t}$, or $\mathbf{P}_X\mathbf{A}_Y\mathbf{t} = \lambda\mathbf{t}$ if a scaling of the PLS regression coefficient in step 3 is assumed. In that case, the procedure is equivalent to canonical correlation analysis applied to $\mathbf{X}$ and $\mathbf{A}_Y$, as is shown below.

The solution for the canonical weight vectors $\mathbf{c}$ (and $\mathbf{q}$) in canonical correlation analysis for two data sets $\mathbf{X}$ and $\mathbf{Y}$, both of full column rank, is obtained from solving the following generalized eigenvalue problem (and a similar one for the Y-weights) [6]:

$$\mathbf{X}^{T}\mathbf{Y}(\mathbf{Y}^{T}\mathbf{Y})^{-1}\mathbf{Y}^{T}\mathbf{X}\mathbf{c} = \lambda\mathbf{X}^{T}\mathbf{X}\mathbf{c} \qquad (2.1)$$

Pre-multiplying both sides by $\mathbf{X}(\mathbf{X}^{T}\mathbf{X})^{-1}$ and writing $\mathbf{t} = \mathbf{X}\mathbf{c}$ leads to

$$\mathbf{X}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{Y}(\mathbf{Y}^{T}\mathbf{Y})^{-1}\mathbf{Y}^{T}\mathbf{t} = \lambda\mathbf{t} \qquad (2.2)$$

or

$$\mathbf{P}_X\mathbf{P}_Y\mathbf{t} = \lambda\mathbf{t} \qquad (2.3)$$

The canonical variate (score) vectors therefore are the eigenvectors of $\mathbf{P}_X\mathbf{P}_Y$. In fact, they are the solution of the symmetrical eigenproblem $\mathbf{P}_X\mathbf{P}_Y\mathbf{P}_X\mathbf{t} = \lambda\mathbf{t}$ since $\mathbf{t} = \mathbf{P}_X\mathbf{t}$. The weights $\mathbf{c}$ follow from regressing $\mathbf{t}$ on $\mathbf{X}$, i.e. $\mathbf{c} = \mathbf{X}^{+}\mathbf{t}$. By interchanging $\mathbf{X}$ and $\mathbf{Y}$ in the above expressions the solution for canonical $\mathbf{Y}$ weights $\mathbf{q}$ and scores $\mathbf{u}$ is obtained. Canonical correlation analysis can therefore be seen as dealing with eigenproblems involving products of orthogonal projection matrices onto the two spaces involved. Hence, Wold's approach for finding OSC directions boils down to a canonical correlation analysis of $\mathbf{X}$ and the orthogonal complement of $\mathbf{Y}$ ($\mathbf{A}_Y$), if all PLS components are chosen to calculate the projection $\mathbf{TT}^{+}$ in step 3, and if $\mathbf{X}$ and $\mathbf{Y}$ are of full column rank.

Only the spaces spanned by the columns of $\mathbf{T}$ and $\mathbf{A}_Y$ play a role, not the (co)variance structures as $\mathbf{P}_T$ and $\mathbf{A}_Y$ carry no information about this. The matrix product $\mathbf{P}_T\mathbf{A}_Y$ only carries information about the inter-correlation structure of $\mathbf{T}$ and $\mathbf{A}_Y$. This means that the OSC method is not implicitly looking for directions that describe large variance in $\mathbf{T}$.

In the last paragraph of their paper, the authors already remark that if $I < J$, there are many solutions. In that case, $\mathbf{P}_T\mathbf{A}_Y$ has multiple eigenvalues equal to 1 and some smaller than 1. This means that there is no dominant eigenvalue and many equivalent solutions can be found. However, the solution found is one close to the starting vector $PC_1(\mathbf{X})$, and therefore this solution describes $\mathbf{X}$ rather well. Furthermore, because of step 3, $\mathbf{t}$ is the PLS fit of $\mathbf{t}^{*}$ to $\mathbf{X}$ and therefore it will also account for a large part of the variation in $\mathbf{X}$.

In the case that $I < J$, $\mathbf{Y}$ will always be in the X-space and $\mathbf{t}^{*}$ will also always be in the X-space. However, $\mathbf{t}$ is likely to be different from $\mathbf{t}^{*}$, due to the PLS regression in step 3, which may cause $\mathbf{t}$ to be correlated to $\mathbf{Y}$. If $I - 1 > J$, $\mathbf{A}_Y\mathbf{t}$ may not be in $\mathbf{X}$, and thus not attainable from $\mathbf{X}$. In that case $\mathbf{t}$ will surely be different from $\mathbf{t}^{*}$ and as a result $\mathbf{t}$ might not be orthogonal to $\mathbf{Y}$. If $\mathbf{t}$ is not orthogonal to $\mathbf{Y}$, information in $\mathbf{X}$ that is relevant to predict $\mathbf{Y}$ will be removed.

2.2. Sjöblom et al.

The algorithm presented by Sjöblom et al. [2], can be described in the following steps

1. $\mathbf{t} = PC_1(\mathbf{X})$
2. $\mathbf{t}^{*} = \mathbf{A}_Y\mathbf{t}$
3. $\mathbf{w} = \mathbf{X}^{T}\mathbf{t}^{*}$, $\mathbf{w} = \mathbf{w}/\|\mathbf{w}\|$
4. $\mathbf{t} = \mathbf{X}\mathbf{w}$

Repeat steps 2–4 until convergence of $\mathbf{t}^*$.

5. Find $\mathbf{t}^{**} = \mathbf{T}\mathbf{T}^{+}\mathbf{t}^*$ using PLS model based on 15 PLS components
6. $\mathbf{p} = \mathbf{X}^{\mathrm{T}}\mathbf{t}^{**}/(\mathbf{t}^{**\mathrm{T}}\mathbf{t}^{**})$
7. $\mathbf{X}^{\text{Sjöblom}} = \mathbf{X} - \mathbf{t}^{**}\mathbf{p}^{\mathrm{T}}$

In step 5, a PLS model with 15 components is calculated between $\mathbf{t}^*$ and $\mathbf{X}$. $\mathbf{t}^*$ is projected on the scores $\mathbf{T}$ of the PLS model in the same way as in the approach by Wold et al. Here, $\mathbf{t}^{**}$ is the score vector of the OSC component. More OSC components can be obtained by repeating the whole procedure starting with $\mathbf{X}^{\text{Sjöblom}}$. For spectra of new samples $\mathbf{x}_{\text{new}}$, the correction is performed by repeating the following steps for each OSC component i:

8. $\mathbf{t}_{\text{new}}^{**} = \mathbf{x}_{\text{new}}^{\mathrm{T}}\mathbf{b}_i$
9. $\mathbf{x}_{\text{new}}^{\text{Sjöblom}} = \mathbf{x}_{\text{new}} - t_{\text{new}}^{**}\mathbf{p}_i$

where $\mathbf{b}_i$ is the PLS regression coefficient of the 15-component PLS model in step 5 for the ith OSC component. Repeated application of steps 2–4 leads to $\lambda\mathbf{t}^* = \mathbf{A}_{\mathrm{Y}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{t}^*$, so this algorithm establishes a sequence of $\mathbf{A}_{\mathrm{Y}}$ and $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ applied to the starting vector. In other words, it finds the dominant eigenvector of $\mathbf{A}_{\mathrm{Y}}\mathbf{X}\mathbf{X}^{\mathrm{T}}$. In this case, a dominant solution exists and thus only one solution is found. Since $\mathbf{t}^*$ lies in the space of $\mathbf{A}_{\mathrm{Y}}$, $\mathbf{t}^*$ is orthogonal to $\mathbf{Y}$ and we may write $\lambda\mathbf{t} = \mathbf{A}_{\mathrm{Y}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{A}_{\mathrm{Y}}\mathbf{t}^*$, which means that $\mathbf{t}^*$ is the first PCA score vector of $\mathbf{A}_{\mathrm{Y}}\mathbf{X}$: $\mathbf{t}^* = \mathrm{PC}_1(\mathbf{A}_{\mathrm{Y}}\mathbf{X})$. Now, $\mathbf{t}^*$ is orthogonal to $\mathbf{Y}$ but may not be in the X-space. Therefore, steps 5 and 6 are applied. Here, a 15-component PLS model is built between $\mathbf{X}$ and $\mathbf{t}^*$ to project $\mathbf{t}^*$ on $\mathbf{T}$; $\mathbf{t}^{**} = \mathbf{P}_{\mathrm{T}}\mathbf{t}^*$, so $\mathbf{t}^{**}$ lies in the X-space, but $\mathbf{t}^{**}$ may be unequal to $\mathbf{t}^*$ and thus $\mathbf{t}^{**}$ may not be orthogonal to $\mathbf{Y}$.

Concluding, $\mathbf{t}^{**} = \mathbf{P}_{\mathrm{T}}(\mathrm{PC}_1(\mathbf{A}_{\mathrm{Y}}\mathbf{X}))$, meaning that $\mathbf{t}^{**}$ is the first PC score vector of $\mathbf{A}_{\mathrm{Y}}\mathbf{X}$, $\mathbf{t}^*$, projected on $\mathbf{T}$, thus the variance described by $\mathbf{t}^*$ is taken into account. This means that the algorithm looks for OSC solutions $\mathbf{t}^*$ that start out to be orthogonal to $\mathbf{Y}$ and account for the largest part of the variance in $\mathbf{X}$. However, in the end, $\mathbf{t}^{**}$ may not be orthogonal to $\mathbf{Y}$.

2.3. Wise and Gallagher

The approach by Wise and Gallagher is basically the same as the approach of Sjöblom et al., but they try to cure for the non-orthogonality problem in Sjöblom's OSC solution. Therefore a last step was added to orthogonalize $\mathbf{t}^{**}$ to $\mathbf{Y}$, $\mathbf{t}^{***} = \mathbf{A}_{\mathrm{Y}}\mathbf{P}_{\mathrm{T}}(\mathrm{PC}_1(\mathbf{A}_{\mathrm{Y}}\mathbf{X}))$. The loading $\mathbf{p}$ then equals, $\mathbf{p} = \mathbf{X}^{\mathrm{T}}\mathbf{t}^{***}/(\mathbf{t}^{***\mathrm{T}}\mathbf{t}^{***})$. The problem with this cure is that if $I - 1 > J$, there is no guarantee that the solution $\mathbf{t}^{***}$ lies in the X-space, and thus removing $\mathbf{t}^{***}$ from $\mathbf{X}$ may introduce components outside the X-space into the corrected matrix. The approaches chosen by Sjöblom et al. and Wise and Gallagher lead to a dilemma, where either the OSC solution is not orthogonal to $\mathbf{Y}$, or it does not lie in the X-space, depending whether the last step is a projection on $\mathbf{X}$ or an anti-projection on $\mathbf{Y}$. For the correction of new spectra of new samples $\mathbf{x}_{\text{new}}$, the following steps are repeated for each OSC component i:

$$t_{\text{new}}^{**} = \mathbf{x}_{\text{new}}^{\mathrm{T}}\mathbf{b}_i$$

$$\mathbf{x}_{\text{new}}^{\text{Wise}} = \mathbf{x}_{\text{new}} - t_{\text{new}}^{**}\mathbf{p}_i$$

where $\mathbf{b}_i$ is the PLS regression coefficient of the many component PLS model in step 5 for the ith OSC component. Note that for new data, t_{new}^{**} is calculated from the PLS model in step 5, but t_{new}^{***} cannot be obtained since $\mathbf{A}_{\mathrm{Y}}t_{\text{new}}^{**}$ is undefined for future samples. The loadings $\mathbf{p}_i$, that correspond to scores t_{new}^{***} are used in combination with t_{new}^{**} to remove the OSC component from $\mathbf{x}_{\text{new}}$. This removal is therefore non-optimal.

2.4. Andersson's direct orthogonalization

Andersson's approach is as follows:

1. $\mathbf{Z} = \mathbf{A}_{\mathrm{Y}}\mathbf{X}$
2. $\mathbf{Z} = \mathbf{t}_{\mathrm{Z}}\mathbf{p}_{\mathrm{Z}}^{\mathrm{T}} + \mathbf{E}$ (using PCA)
3. $\mathbf{t}_{\mathrm{X}} = \mathbf{X}\mathbf{p}_{\mathrm{Z}}$
4. $\mathbf{X}^{\mathrm{DO}} = \mathbf{X} - \mathbf{t}_{\mathrm{X}}\mathbf{p}_{\mathrm{Z}}^{\mathrm{T}}$

If more OSC components are needed, this can be adjusted in the PCA in step 2. For spectra of new samples $\mathbf{x}_{new}$, the correction is performed as follows:

5. $t_{new} = \mathbf{x}_{new}^{T}\mathbf{p}_{Z}$
6. $\mathbf{x}_{new}^{DO} = \mathbf{x}_{new} - \mathbf{t}_{new}\mathbf{p}_{Z}$

In step 2, PC($\mathbf{A}_{Y}\mathbf{X}$) provides the score $\mathbf{t}_{Z}$ and loading $\mathbf{p}_{Z}$. However, $\mathbf{A}_{Y}\mathbf{X}$ may not lie in the X-space when $I-1>J$ and thus the loadings do not necessarily describe a direction in $\mathbf{X}$. In step 3, $\mathbf{X}$ is projected on the loadings of $\mathbf{A}_{Y}\mathbf{X}$ providing $\mathbf{t}_{X}$. However, $\mathbf{t}_{X}$ has no clear distinct properties. It does not describe the largest variance in $\mathbf{X}$ orthogonal to $\mathbf{Y}$, and it also may not be orthogonal to $\mathbf{Y}$. Furthermore, the deflation in step 4 is not optimal because loadings $\mathbf{p}_{Z}$ are obtained from the PCA on $\mathbf{Z}$ instead of $\mathbf{X}$. This might be improved with using $\mathbf{p}_{X}$ instead of $\mathbf{p}_{Z}$, where $\mathbf{p}_{X} = \mathbf{X}^{T}\mathbf{t}_{X}(\mathbf{t}_{X}^{T}\mathbf{t}_{X})^{-1}$. This, however, does not solve the problem of non-orthogonality.

2.5. Fearn

Fearn [4] poses the OSC problem as follows:

1. $\mathbf{t} = \mathbf{X}\mathbf{r}$
2. $\max(\mathbf{t}^{T}\mathbf{t})$, subject to $\mathbf{r}^{T}\mathbf{r} = 1$ and $\mathbf{t}^{T}\mathbf{Y} = 0$.

The solution of $\mathbf{r}$ is obtained as the eigenvector of $\mathbf{A}_{\mathbf{X}^{T}\mathbf{Y}}\mathbf{X}^{T}\mathbf{X}$ corresponding to the largest eigenvalue. Thus, $\mathbf{t} = \mathbf{X}\mathrm{PC}_{1}(\mathbf{A}_{\mathbf{X}^{T}\mathbf{Y}}\mathbf{X}^{T}\mathbf{X})$, which clearly lies in the X-space.

3. $\mathbf{p} = \mathbf{X}^{T}\mathbf{t}/(\mathbf{t}^{T}\mathbf{t})$
4. $\mathbf{X}^{\mathrm{Fearn}} = \mathbf{X} - \mathbf{t}\mathbf{p}^{T}$

A second OSC component can be calculated using the same steps and starting with $\mathbf{X}^{\mathrm{Fearn}}$ as $\mathbf{X}$. For spectra of new samples, $\mathbf{x}_{new}$, the correction is performed by repeating the following steps for each of the i OSC components:

5. $t_{new} = \mathbf{x}_{new}^{T}\mathbf{r}_{i}$
6. $\mathbf{x}_{new}^{\mathrm{Fearn}} = \mathbf{x}_{new} - \mathbf{t}_{new}\mathbf{p}_{i}$

In this approach, a score vector $\mathbf{t}$ in X-space is found that is orthogonal to $\mathbf{Y}$ and *has* maximum variance (subject to the weight vector $\mathbf{r}$ having unit norm). Due to the orthogonality constraint, this direction is usually different from the one that *accounts for* the largest variance in $\mathbf{X}$. In Appendix A, the relation between Fearn's approach and DOSC is presented.

2.6. Problems with OSC

In the latent variable methods such as PLS or PCR, both the measurement error in $\mathbf{Y}$ and in $\mathbf{X}$ are taken into account to come up with a robust calibration model. However, in OSC the measurement error in $\mathbf{Y}$ is fully disregarded because absolute orthogonality is demanded for the OSC components in $\mathbf{X}$. This conflict has manifested itself in such a way that complete orthogonality is not obtained by some of the methods presented in literature (Wold et al. and Sjöblom et al., Andersson) or the final OSC component does not lie in the X-space (Wise and Gallagher, Andersson). The latter introduces new components outside the X-space into the corrected matrix when the OSC component is deflated. Fearn's method, which uses only least squares steps does not have these problems, but is suboptimal in describing the maximum variance of $\mathbf{X}$ with the OSC component.

3. Direct orthogonal signal correction (DOSC)

The direct orthogonal signal correction (DOSC) approach is solely based on least squares steps. It will always find components, which are orthogonal to $\mathbf{Y}$, that describe the largest variation of $\mathbf{X}$.

The first step of DOSC is to decompose $\mathbf{Y}$ into two orthogonal parts, the projection of $\mathbf{Y}$ onto $\mathbf{X}$, $\hat{\mathbf{Y}}$, and the residual part, $\mathbf{F}$ that is orthogonal to $\mathbf{X}$:

1. $\mathbf{Y} = \mathbf{P}_{X}\mathbf{Y} + \mathbf{A}_{X}\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{F}$

Next, $\mathbf{X}$ is decomposed into two orthogonal parts, one part that has the same range as $\hat{\mathbf{Y}}$ and another part that is orthogonal to it:

2. $\mathbf{X} = \mathbf{P}_{\hat{Y}}\mathbf{X} + \mathbf{A}_{\hat{Y}}\mathbf{X}$

Note that for spectral data commonly $J > I$, in which case $\hat{\mathbf{Y}} = \mathbf{Y}$ and then $\mathbf{X}$ may be orthogonalized directly with respect to observed $\mathbf{Y}$, as in step 3.

3. $\mathbf{X} = \mathbf{P}_{\mathrm{Y}}\mathbf{X} + \mathbf{A}_{\mathrm{Y}}\mathbf{X}$ for $J > I$

For non-spectral data where $J < I$, however, it is essential to project $\mathbf{X}$ on $\hat{\mathbf{Y}}$ rather than on $\mathbf{Y}$, since $\hat{\mathbf{Y}} = \mathbf{P}_{\mathrm{X}}\mathbf{Y}$ is in the range of $\mathbf{X}$, and so is $\mathbf{A}_{\hat{\mathrm{Y}}}\mathbf{X} = \mathbf{X} - \mathbf{P}_{\hat{\mathrm{Y}}}\mathbf{X}$. The columns of $\mathbf{A}_{\hat{\mathrm{Y}}}\mathbf{X}$ therefore span a subspace of $\mathbf{X}$ that is orthogonal both to $\hat{\mathbf{Y}}$ and to $\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{F}$, since $\mathbf{F}$ is also orthogonal to $\mathbf{X}$.

Having found this orthogonal subspace $\mathbf{A}_{\hat{\mathrm{Y}}}\mathbf{X}$, PCA is now applied to find the principal component $\mathbf{t}$ corresponding to the largest singular value. If more DOSC components are necessary, more principal components can be obtained in this step. $\mathbf{t}$ is a basis for the one-dimensional subspace that accounts for maximum variance of $\mathbf{A}_{\hat{\mathrm{Y}}}\mathbf{X}$. This then is the sought for one-dimensional subspace of $\mathbf{X}$ that is orthogonal to $\mathbf{Y}$ *and* accounts for the maximum possible variance of $\mathbf{X}$. We finally express the directions $\mathbf{t}$ as linear combinations of $\mathbf{X}$:

4. $\mathbf{t} = \mathbf{X}\mathbf{r}$

with

5. $\mathbf{r} = \mathbf{X}^{+}\mathbf{t}$

where $\mathbf{X}^{+}$ is the Moore–Penrose inverse of $\mathbf{X}$. The large-variance zero-correlation part of $\mathbf{X}$ that we do not use in subsequent regression modeling is removed from the data:

6. $\mathbf{X}^{\mathrm{DOSC}} = \mathbf{X} - \mathbf{P}_{\mathrm{t}}\mathbf{X} = \mathbf{X} - \mathbf{t}(\mathbf{t}^{\mathrm{T}}\mathbf{t})^{-1}\mathbf{t}^{\mathrm{T}}\mathbf{X} = \mathbf{X} - \mathbf{t}\mathbf{p}^{\mathrm{T}} = \mathbf{X} - \mathbf{X}\mathbf{r}\mathbf{p}^{\mathrm{T}}$

with

7. $\mathbf{p} = \mathbf{X}^{\mathrm{T}}\mathbf{t}(\mathbf{t}^{\mathrm{T}}\mathbf{t})^{-1}$

For spectra of new samples $\mathbf{x}_{\mathrm{new}}$, the correction can be performed as follows:

8. $\mathbf{x}_{\mathrm{new}}^{\mathrm{DOSC}} = \mathbf{x}_{\mathrm{new}} - \mathbf{r}^{\mathrm{T}}\mathbf{x}_{\mathrm{new}}\mathbf{p}$

Now $\mathbf{x}_{\mathrm{new}}^{\mathrm{DOSC}}$ can be used in the calibration model instead of $\mathbf{x}_{\mathrm{new}}$ to predict $\mathbf{y}_{\mathrm{new}}$.

Note that in step 5, in order to calculate the weight vector $\mathbf{r}$, the Moore–Penrose inverse $\mathbf{X}^{+}$ is used. This specific inverse is exact meaning that $\mathbf{t}$ exactly equals $\mathbf{X}\mathbf{r}$. The DOSC approach was implemented and tested on some spectral data sets, as discussed in Section 4. However, our first practical results with the new method were not encouraging. If the spectra are corrected according to the DOSC method, the test set predictions of $\mathbf{Y}$ are worse than if no orthogonal signal correction is used. This is probably due to the constraint of complete orthogonality. Even the non-stable directions in $\mathbf{X}$ are used to fit the DOSC component $\mathbf{t}$. This leads to an overfit of this DOSC component. The problem of overfit can be solved by loosening the complete orthogonality constraint. The exact fit of $\mathbf{t}$ using the Moore–Penrose inverse $\mathbf{X}^{+}$ in step 5 is loosened by using a generalized inverse $\mathbf{X}^{-}$ which is not completely exact, i.e. $\mathbf{t} \approx \tilde{\mathbf{t}} = \mathbf{X}\mathbf{r}$ (see Appendix B). The generalized inverse $\mathbf{X}^{-}$ is calculated using a PCR solution between $\mathbf{X}$ and $\mathbf{t}$. In this case only the stable directions in $\mathbf{X}$ are used to define $\tilde{\mathbf{t}}$. The number of principal components for the PCR solution equals the number of singular values of $\mathbf{X}$ larger than a tolerance factor, which has to be tuned.

5a. $\tilde{\mathbf{r}} = \mathbf{X}^{-}\mathbf{t}$, $\mathbf{t} = \mathbf{X}\tilde{\mathbf{r}} + \mathbf{e}$, $\tilde{\mathbf{t}} = \mathbf{X}\tilde{\mathbf{r}}$

This leads to

6a. $\mathbf{X}^{\mathrm{DOSC}} = \mathbf{X} - \tilde{\mathbf{t}}\tilde{\mathbf{p}}^{\mathrm{T}} = \mathbf{X} - \mathbf{X}\tilde{\mathbf{r}}\tilde{\mathbf{p}}^{\mathrm{T}}$

with

7a. $\tilde{\mathbf{p}} = \mathbf{X}^{\mathrm{T}}\tilde{\mathbf{t}}(\tilde{\mathbf{t}}^{\mathrm{T}}\tilde{\mathbf{t}})^{-1}$

In step 5a, $\tilde{\mathbf{t}}$ is different from $\mathbf{t}$ and is allowed to go slightly out of the $\mathbf{A}_{\hat{\mathrm{Y}}}\mathbf{X}$ space. Fig. 1a shows schematically the effect of loosening the orthogonality constraint of the DOSC component. In this figure, the multivariate space defined by the range of $\mathbf{X}$ is presented together with the projection of $\mathbf{Y}$, $\hat{\mathbf{Y}}$, in this space. The first principal component of the space orthogonal to $\hat{\mathbf{Y}}$, which is the DOSC component $\mathbf{t}$, will

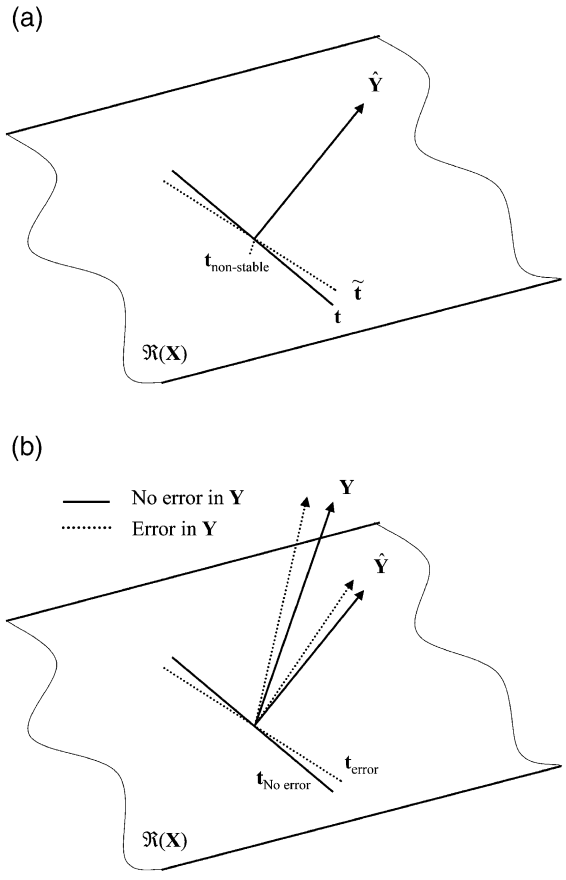

Fig. 1. (a) Effect of using only the stable directions of **X** to determine the DOSC component. (b) Effect of measurement error in **Y** on final DOSC component.

have contributions in all directions of **X**, including the non-stable directions. Using a PCR-type solution to fit **t**, by means of using the generalized inverse $\mathbf{X}^{-}$, in step 5a, only the stable directions in **X** are used. The part of **t** that comes from the non-stable directions in **X**, $\mathbf{t}_{\text{non-stable}}$, is removed. This leaves a direction $\tilde{\mathbf{t}}$ which is fitted using only the stable directions in **X**. However, $\tilde{\mathbf{t}}$ lost its property of complete orthogonality with $\hat{\mathbf{Y}}$.

Another possible reason why rotating **t** towards $\tilde{\mathbf{t}}$ leads to better predictions for new data is that the complete orthogonality constraint disregards the measurement error in **Y**. Different from the PLS approach where latent variables are used to deal with measurement error in both **X** and **Y**, the OSC approach asks for exact orthogonality with **Y**. In Fig. 1b, the same multivariate space is presented as in Fig. 1a, together with both **Y** and $\hat{\mathbf{Y}}$. If **Y** was known without measurement error (solid line), the DOSC procedure would lead to the DOSC component $\mathbf{t}_{\text{No error}}$. This component will still slightly be present in the non-stable directions, but it will be optimally defined in the stable directions. Due to measurement error in **Y** (dotted line), a slightly different DOSC component, $\mathbf{t}_{\text{error}}$, is found. This component will be non-optimally defined in the stable directions in **X**. However, due to the rotation of $\mathbf{t}_{\text{error}}$ to $\tilde{\mathbf{t}}$ as described above, the component is allowed to change to more robust directions of **X**.

A combination of the effects described above will be the reason for the improved test set predictions as will be presented in Section 4. However, this problem is still not completely understood and further research is necessary.

A different approach to calculate a direct orthogonal signal correction was developed at the same time. This approach, however, turned out to be equal to DOSC. Appendix A shows this approach and the proof for equality. This alternative approach is shown for a better understanding of the problem and its solution.

3.1. Tuning of DOSC

A problem in applying any orthogonal signal correction method is to optimally tune the system. First of all, for each calibration model the optimal number of OSC components and the optimal number of PLS components have to be determined. Furthermore, there is another meta-parameter to tune. This parameter is in the approaches of Wold et al., Sjöblom et al., and Wise and Gallagher, the number of PLS components used to calculate each OSC component (see step 3 of Wold's approach and step 5 in Sjöblom et al. and Wise and Gallagher). Although Wise and Gallagher and Sjöblom et al. both give default values for this number, it can be tuned to minimize prediction errors. In DOSC, the optimal generalized inverse, using the optimal number of PCR components, has to be picked. It does not seem a good idea to do a full cross validation to find optimal values for each of the meta-parameters. Wold et al. already started some discussion on the number of OSC com-

ponents to choose. The number of OSC components should not be too high, since this might lead to overfit of the model. One or two OSC components are usually sufficient. The first OSC component often resembles a base-line correction and the second can correct for multiplicative effects [1].

4. Results and discussion

In this section, two applications will be presented where different types of OSC methods are used to remove parts of the spectral data to improve the calibration model. Improvement of calibration models means that the prediction is improved or that the model needs less latent variables to obtain the same prediction quality.

4.1. Prediction of viscosity of diesel fuels

The first application deals with the NIR spectra of diesel fuels. These spectra have been measured at Southwest Research Institute (SWRI) on a project sponsored by the U.S. Army. The data were obtained from the Eigenvector Research homepage (www.eigenvector.com). The training set consists of NIR spectra of 136 diesel fuels. The viscosity of the diesel fuels was obtained using a separate measurement. The test set consists of 116 diesel fuels.

A PLS calibration model was developed between the mean centered NIR spectra of the training set of diesel fuels and the viscosity of the fuels. A large number of components were necessary for a good prediction of the viscosity. Fig. 2 shows the root mean squared error of prediction (RMSEP) of the viscosity of the fuels of the test set versus the number of PLS components of the calibration model. After 12 PLS components, the RMSEP did not decrease further. In order to decrease the large number of components, OSC components were calculated and removed from the spectral data. In this case for all methods, only one OSC component is calculated, because the aim of this paper is to show the differences between the OSC components of the different methods, and not to minimize the prediction error of the product quality. Table 1 shows the results obtained for the different OSC methods. The percentage of variation removed from the training set as well as from the test set are given, and the correlation of the OSC component with the **Y** variable is given. This correlation is presented to show that some methods remove parts of **X** that are correlated with **Y**.

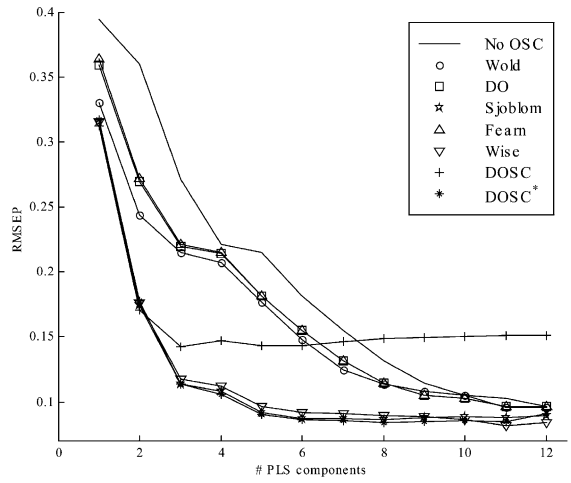

Fig. 2. RMSEP values of the viscosity prediction of diesel fuel samples using PLS models with 1–12 components after one OSC component has been removed using different OSC methods.

The results in Table 1, show that DO and the approach of Sjöblom et al., give OSC components that are somewhat correlated with **Y**. The percentage of variation of **X** removed is the highest for DO and OSC Sjöblom, but this is probably due to the fact that the orthogonal constraint is not fully applied. The solution given by Fearn clearly removes the lowest percentage of variation.

Using only one OSC component, the minimum RMSEP is obtained at a lower number of components compared to the case when no OSC is used. The RMSEP profiles obtained using the Wold, DO or the Fearn approaches, follow the RMSEP profile of the standard PLS model without using OSC, except the RMSEP values are shifted one PLS component to the left. This means that one PLS component is replaced by one OSC component. The total number of components remains equal, which means that the models are not simplified. The gain using the Wise and Sjöblom approaches is much larger. The minimum RMSEP, which is even lower than the minimum RMSEP without using OSC, is already obtained after six PLS components. The standard DOSC approach with the Moore–Penrose inverse has some prediction prob-

Table 1
The results of the different OSC transformations using only one OSC component for the diesel fuel data

Method	% Removed from training set	Correlation of OSC component with **Y**	% Removed test set
OSC Wold	89.30	2×10^{-7}	90.07
DO	90.20	8×10^{-2}	90.87
OSC Wise	89.55	0	90.11
OSC Sjöblom	89.60	3×10^{-3}	90.10
Fearn	88.12	0	88.72
DOSC	89.58	0	89.90
DOSC*	89.60	3×10^{-3}	90.14

lems. The RMSEP only drops to 0.14, which is much higher than the minimum RMSEP for all other methods. If, instead of the Moore–Penrose inverse, another generalized inverse is used (here the tolerance for DOSC* is set to 1×10^{-3}) then the RMSEP is even lower than that of the Wise and Sjöblom approaches. This DOSC* method, however, loosens the orthogonality constraint which leads to a correlation of 0.003 with between the DOSC* component and **Y**. The results of the Sjöblom approach and DOSC* are very similar. Both methods have the same correlation with **Y**, the variation removed in both the training and the test set are similar as well as the RMSEP values of the prediction of the viscosity of the test set samples.

The tolerance factor used to calculate the generalized inverse of **X** in the DOSC approach is a critical value. It determines the number of singular vectors, or PCR components of **X**, that are used in calculating the generalized inverse in step 5a of the DOSC approach. If a generalized inverse $\mathbf{X}^-$ is used to calculate $\tilde{\mathbf{r}}$, $\tilde{\mathbf{t}}$ is unequal to **t**. This also means that $\tilde{\mathbf{t}}$ is not necessarily orthogonal to **Y**. The results in Table 1 and Fig. 2 were obtained with a tolerance factor of 10^{-3}. In that case 24 singular vectors and values were used to calculate the inverse. Table 2 shows some results when the tolerance factor is changed.

If the tolerance is increased from 10^{-6} to 5×10^{-2}, the number of singular values used to calculate the inverse decreases from 135 to 2. Here, the generalized inverse $\mathbf{X}^-$ with tolerance of 10^{-6} equals the Moore–Penrose (MP) inverse $\mathbf{X}^+$. If the tolerance is increased, then $\tilde{\mathbf{t}}$ slowly moves away from **t**. The squared correlation, between **t** and $\tilde{\mathbf{t}}$ decreases. This causes the correlation between $\tilde{\mathbf{t}}$ and **Y** to increase up to 0.07. The percentage variation removed by the OSC component increases, but this is because the orthogonality constraint is loosened.

Fig. 3 shows the RMSEP of the test set for different values of the tolerance. A tolerance of 10^{-3} in this case is clearly the optimal value, however for new data sets the tolerance factor should be tuned properly. This means that allowing some correlation between the OSC component and **Y** gives better predictions for a separate test set.

The methods of Wold et al., Sjöblom et al., and Wise and Gallagher all use a PLS model of **X** to calculate PLS regression coefficients **b** (see Section 2.1 step 3 and Section 2.2 step 5). This **b** is obtained by calculating a many-component PLS model between **X** and $\mathbf{t}^*$, to make sure that the solution is in the X-space and that the solution accounts for a large part of the variation in **X**. Sjöblom et al. use 15 components for the PLS model, and Wise and Gallagher want to describe at least 99.9% of the variation in **X** with the PLS model (which is the default setting). For this specific data set this comes down to use 13 PLS

Table 2
The effect of tolerance in calculating the generalized inverse of **X** on the correlation with **Y** and the percentage of **X** explained of the diesel fuel data

Tolerance	% Removed from training set	Correlation of OSC component with **Y**	# Singular vectors and values used to calculate inverse	Squared correlation between $\tilde{\mathbf{t}}$ and **t**	% Removed from test set
MP	89.58	0	135	1.0000	89.90
1×10^{-4}	89.59	1×10^{-3}	67	0.9999	89.98
1×10^{-3}	89.60	3×10^{-3}	24	0.9997	90.14
1×10^{-2}	89.73	2×10^{-2}	7	0.9982	90.30
5×10^{-2}	90.13	7×10^{-2}	2	0.9938	90.81

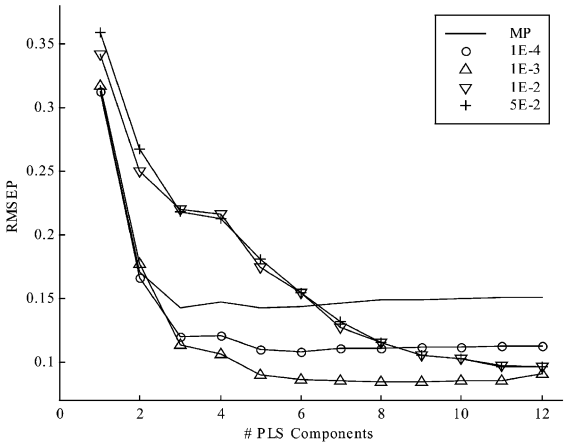

Fig. 3. RMSEP values of the viscosity prediction of diesel fuel samples using PLS models with 1–12 components after one OSC component has been removed using DOSC with different tolerance factors.

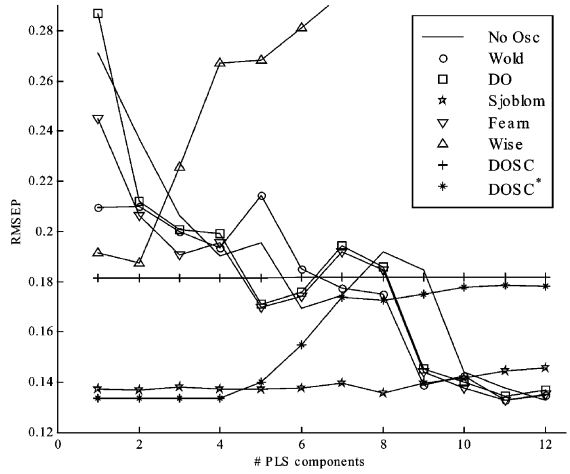

Fig. 4. RMSEP values of the moisture prediction of corn samples using PLS models with 1–12 components after one OSC component has been removed using different OSC methods.

components. Wold et al. do not describe precisely how many components are used to calculate the PLS inverse. For the results obtained in this paper, 10 PLS components were used.

The DOSC* approach with tolerance 10^{-3} uses 24 singular vectors or PCR components of $\mathbf{X}$ to calculate the generalized inverse $\mathbf{X}^-$. This is somewhat more than Sjöblom et al. and Wise and Gallagher use. However, the correlation of the OSC component and $\mathbf{Y}$ is comparable to the results of Sjöblom et al., and also the percentages explained in both training set and test set are comparable.

4.2. Prediction of moisture content of corn samples

For the second application, NIR spectra of corn samples were obtained to predict the moisture content of the corn. These spectra were also obtained from the Eigenvector Research homepage. The wavelength range is 1100–2498 nm at 2-nm intervals (700 channels). The moisture, oil protein and starch values are measured for each of the samples. The data was originally taken at Cargill.

Spectra of 80 corn samples are available. These were divided into a training set (47 samples) and a test set (31 samples), and two samples were rejected as outliers. The spectra used in this example were obtained from spectrometer mp5 and the moisture in the corn samples was used as the response variable. A PLS model between the mean centered spectra and the moisture of the corn samples was built. Fig. 4 shows the RMSEP values for the prediction of the moisture for the corn samples in the test set when one OSC component was removed from the spectra. Without OSC, the minimum RMSEP value of 0.13 is obtained for a PLS model with 10 components. If one OSC component is removed using the approaches of Wold, Fearn and the DO approach, the same RMSEP level is obtained with only a nine-component PLS model. Again the calibration model is not simplified because the total number of components is still 10.

Table 3
The results of the different OSC transformations using only one OSC component for the corn data

Method	% Removed from training set	Correlation of OSC component with $\mathbf{Y}$	% Removed test set
OSC Wold	59.74	5×10^{-8}	54.40
DO	98.96	6×10^{-1}	99.12
OSC Wise	58.90	0	57.39
OSC Sjöblom	65.75	2×10^{-2}	69.77
Fearn	31.64	0	35.24
DOSC	64.98	0	69.14
DOSC*	65.63	1×10^{-2}	70.06

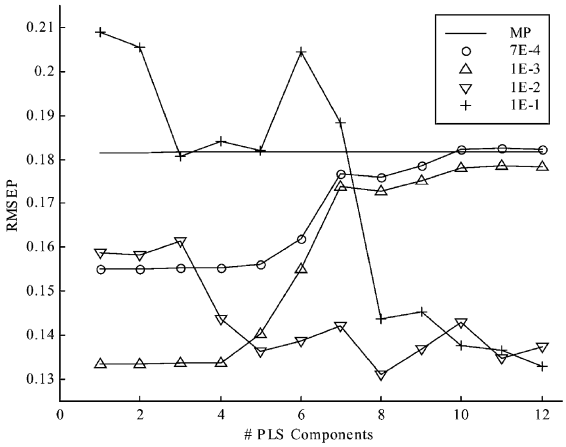

Fig. 5. RMSEP values of the moisture prediction of corn samples using PLS models with 1–12 components after one OSC component has been removed using DOSC with different tolerance factors.

The approach of Wise does not work well in this example, but that may be improved by tuning it properly. The standard DOSC approach with a Moore–Penrose gives an RMSEP of 0.18, but if it is tuned properly with a tolerance of 10^{-3} (DOSC*) the minimum RMSEP level of 0.13 is already obtained with a one-component PLS model. Again the approach of Sjöblom gives results similar to DOSC*, for the variance removed in both training and test set, the correlation with **Y** and the RMSEP values for the prediction of the moisture content. These results are presented in Table 3.

Fig. 5 shows the RMSEP values of the moisture prediction in corn samples when the DOSC approach is tuned. The minimum RMSEP value of 0.13 is obtained for a tolerance value of 10^{-3} or higher. However if the tolerance is higher than 10^{-3}, more PLS components are needed to reach this minimum value. Again the optimal tolerance is found to be near 10^{-3}. Table 4 shows some results for the different tolerance values. The results are similar to the ones obtained in the previous application. If the tolerance is increased, the squared correlation between **t** and $\tilde{\mathbf{t}}$ decreases. This results in an increasing correlation between the DOSC component and the response variable. However, allowing a small correlation improves the prediction of the moisture content of the corn samples and it also reduced the number of PLS components necessary in the calibration model.

4.3. The orthogonality constraint

The orthogonality constraint, which is the basis for the OSC approach is loosened in some methods. If the OSC component is correlated with **Y**, information that can be relevant to predict **Y** is removed from **X**. This will probably lead to a lower fit of **Y** in the resulting PLS model. However, for the prediction of **Y** for new samples this is not necessarily true. In the results presented above it is shown that if a small correlation between the DOSC* component and **Y** is allowed, the prediction error for new samples decreased considerably. The reason for this effect will be a combination between not using the non-stable directions in **X** and taking the measurement error of **Y** into account, as described in more detail at the end of Section 3.

Table 4
The effect of tolerance in calculating the generalized inverse of **X** on the correlation with **Y** and the percentage of **X** explained of the corn data

Tolerance	% Removed from training set	Correlation of OSC component with **Y**	# Singular vectors and values used to calculate inverse	Squared correlation between $\tilde{\mathbf{t}}$ and **t**	% Removed from test set
MP	64.98	0	46	1.0000	69.14
7×10^{-4}	65.32	7×10^{-3}	42	0.9949	69.15
1×10^{-3}	65.63	1×10^{-2}	35	0.9901	70.05
1×10^{-2}	67.04	4×10^{-2}	12	0.9683	69.13
1×10^{-1}	73.95	2×10^{-1}	3	0.8621	56.09

5. Conclusion

In this paper the OSC approach is discussed and five methods are compared. Since all of these methods had some problems, a new direct orthogonal signal correction method, DOSC, was introduced. DOSC calculates components that are orthogonal to **Y** and describe the largest variation in **X**. The method is developed using only simple least squares steps. However, it became clear that a complete orthogonality constraint is too strict, since this leads to the use of non-stable directions in **X**. Furthermore the measurement error in **Y** is disregarded. This leads to an overfit of the DOSC component and therefore predictions for a separate test set can become worse than without using DOSC. One approach to loosen the complete orthogonality constraint is presented and this improves the method considerably.

Two applications are shown where DOSC is used. In both cases the optimal number of PLS components in the final calibration model was reduced considerably by using only one DOSC component. This means that the total calibration model is simplified because the total number of components used is decreased. This was not the case for all OSC methods discussed in this paper. Large reductions of prediction error however, were not observed for any of the methods.

Acknowledgements

The authors want to thank Scott Hutzler of Southwest Research Institute, Mike Blackburn at Cargill and Barry Wise of Eigenvector Research for providing the diesel fuel and corn data.

Appendix A

A.1

Development of DOSC from another starting point. The alternative DOSC selects a direction **t** in **X**, orthogonal to **Y**, that describes the largest variation in **X**. This direction will be removed from **X**. The following minimization has to be solved.

$$\min_{t,p} \|\mathbf{X} - \mathbf{t}\mathbf{p}^T\|^2, \quad \text{s.t.} \quad \{\mathbf{Y}^T\mathbf{t} = 0\},$$

or

$$\min_{w,p} \|\mathbf{X} - \mathbf{X}\mathbf{w}\mathbf{p}^T\|^2, \quad \text{s.t.} \quad \{\mathbf{Y}^T\mathbf{X}\mathbf{w} = 0\} \qquad (A.1)$$

Firstly, it will be proven that $\mathbf{Y}^T\mathbf{X}\mathbf{w}$ equals $\hat{\mathbf{Y}}^T\mathbf{X}\mathbf{w}$.

$$\hat{\mathbf{Y}}^T\mathbf{X}\mathbf{w} = (\mathbf{X}\mathbf{X}^+\mathbf{Y})^T\mathbf{X}\mathbf{w} = \mathbf{Y}^T\mathbf{X}\mathbf{X}^+\mathbf{X}\mathbf{w} = \mathbf{Y}^T\mathbf{X}\mathbf{w}$$

Thus problem (A.1) can also be written as:

$$\min_{w,p} \|\mathbf{X} - \mathbf{X}\mathbf{w}\mathbf{p}^T\|^2, \quad \text{s.t.} \quad \{\hat{\mathbf{Y}}^T\mathbf{X}\mathbf{w} = 0\} \qquad (A.2)$$

The constraint on the direction $\mathbf{X}\mathbf{w}$ to be orthogonal to $\hat{\mathbf{Y}}$ can be expressed directly by forcing the direction to come from the part in **X** orthogonal to $\hat{\mathbf{Y}}$, i.e. from $\mathbf{A}_{\hat{Y}}\mathbf{X}\mathbf{w}$. $\hat{\mathbf{Y}}^T\mathbf{A}_{\hat{Y}}$ equals 0 by definition.

Thus, problem (A.2) can be written as:

$$\min_{w,p} \|\mathbf{X} - \mathbf{A}_{\hat{Y}}\mathbf{X}\mathbf{w}\mathbf{p}^T\|^2 \qquad (A.3)$$

In step 2 of the DOSC (Section 3), **X** is divided into two orthogonal parts, $\mathbf{A}_{\hat{Y}}\mathbf{X}$ and $\mathbf{P}_{\hat{Y}}\mathbf{X}$. Since the direction $\mathbf{A}_{\hat{Y}}\mathbf{X}\mathbf{w}$ lies in the range of $\mathbf{A}_{\hat{Y}}\mathbf{X}$ it can never describe variation in $\mathbf{P}_{\hat{Y}}\mathbf{X}$. Therefore problem (A.3) is equivalent to solving

$$\min_{w,p} \|\mathbf{A}_{\hat{Y}}\mathbf{X} - \mathbf{A}_{\hat{Y}}\mathbf{X}\mathbf{w}\mathbf{p}^T\|^2 \qquad (A.4)$$

which equals the DOSC approach described in Section 3.

A.2. Relation of DOSC with Fearn's approach

Consider the following minimization problem

$$\min_{w} \|\mathbf{X} - \mathbf{X}\mathbf{w}\mathbf{w}^T\|^2, \quad \text{s.t.}$$

$$\{\mathbf{Y}^T\mathbf{X}\mathbf{w} = 0, \mathbf{w}^T\mathbf{w} = 1\} \qquad (A.5)$$

This solution to this problem can be rewritten as follows, where in each step the same constraints are applied as in equation (A.5):

$$\begin{aligned}
&\min_{\mathbf{w}} \left\{ \mathrm{tr}(\mathbf{X} - \mathbf{X}\mathbf{w}\mathbf{w}^{\mathrm{T}})^{\mathrm{T}} (\mathbf{X} - \mathbf{X}\mathbf{w}\mathbf{w}^{\mathrm{T}}) \right\} \\
&\quad = \min_{\mathbf{w}} \left\{ \mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) - \mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}\mathbf{w}^{\mathrm{T}}) \right. \\
&\qquad \left. - \mathrm{tr}(\mathbf{w}\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}) + \mathrm{tr}(\mathbf{w}\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}\mathbf{w}^{\mathrm{T}}) \right\} \\
&\quad = \min_{\mathbf{w}} \left\{ \mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) - 2\mathrm{tr}(\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}) \right. \\
&\qquad \left. - \mathrm{tr}(\mathbf{w}^{\mathrm{T}}\mathbf{w}\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}) \right\} \\
&\quad = \min_{\mathbf{w}} \left\{ \mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) - \mathrm{tr}(\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}) \right\} \\
&\quad = \max_{\mathbf{w}} \left\{ \mathrm{tr}(\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}), \quad \text{s.t.} \right. \\
&\qquad \left. \{\mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{w} = 0, \mathbf{w}^{\mathrm{T}}\mathbf{w} = 1\} \right\}
\end{aligned}$$

The latter formulation is exactly the one introduced by Fearn. Conversely, Fearn's approach can be viewed as solving problem (A.5). Following, the same reasoning as in Appendix A.1 we may rewrite (A.5) as follows:

$$\min_{w} \|\mathbf{A}_{\hat{\mathbf{Y}}}\mathbf{X} - \mathbf{A}_{\hat{\mathbf{Y}}}\mathbf{X}\mathbf{w}\mathbf{w}^{\mathrm{T}}\|^2, \quad \text{s.t.}\, \mathbf{w}^{\mathrm{T}}\mathbf{w} = 1 \qquad (A.6)$$

The problem in Eq. (A.6) equals the DOSC problem (A.3) with two additional restrictions being $\mathbf{p} = \mathbf{w}$ and $|\mathbf{w}| = 1$. Therefore the variation described by the OSC solution given by Fearn will at best be equal to DOSC, but in most cases it will be lower.

Without the orthogonality constraint, problem (A.5) would equal a PCA on **X**. With the constraint active, the method finds the direction with the maximum variance, subject to the weight vector **w** having norm 1, in **X** orthogonal to **Y**. Due to the constraint, in many cases this direction will be different from the direction that accounts for the largest variation in **X**.

Appendix B

The generalized inverse $\mathbf{X}^{-}$, can be calculated using the singular value decomposition of **X**. If

$$\mathbf{X} = \mathbf{U}\begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix}\mathbf{V}^{\mathrm{T}} \qquad (B.1)$$

then

$$\mathbf{X}^{-} = \mathbf{V}\begin{bmatrix} \Delta^{-1} & \mathbf{E} \\ \mathbf{F} & \mathbf{G} \end{bmatrix}\mathbf{U}^{\mathrm{T}} \qquad (B.2)$$

is a generalized inverse of **X** for all choices of **E**, **F** and **G** with the correct sizes [7]. In this application **E**, **F** and **G** are set to zero, and the singular values smaller than the tolerance value are set to zero in Δ. This generalized inverse corresponds to a PCA solution of **X**, where the number of principal components used equals the number of singular values larger than the tolerance factor. If all nonzero singular values in Δ are used to calculate the generalized inverse, then this equals the Moore–Penrose inverse.

References

[1] S. Wold, H. Antti, F. Lindgren, J. Ohman, Orthogonal signal correction of near-infrared spectra. Chemometrics and Intelligent Laboratory Systems 44 (1998) 175–185.

[2] J. Sjöblom, O. Svensson, M. Josefson, H. Kullberg, S. Wold, An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. Chemometrics and Intelligent Laboratory Systems 44 (1998) 229–244.

[3] C.A. Andersson, Direct orthogonalization. Chemometrics and Intelligent Laboratory Systems 47 (1999) 51–63.

[4] T. Fearn, On orthogonal signal correction. Chemometrics and Intelligent Laboratory Systems 50 (2000) 47–52.

[5] B.M. Wise, N.B. Gallagher, http://www.eigenvector.com/MATLAB/OSC.html.

[6] R. Gittins, Canonical Analysis. A Review with Applications in Ecology 1985 Berlin.

[7] J.R. Schott, Matrix Analysis for Statistics. Wiley, New York, 1997 .