A didactically motivated PLS prediction algorithm

ROLF ERGON and KIM H. ESBENSEN

Keywords: *Latent variables models, PLS prediction, Kalman filtering*

The intention of this paper is to develop an easily understood PLS prediction algorithm, especially for the control community. The algorithm is based on an explicit latent variables model, and is otherwise a combination of the previously published Martens and Helland algorithms. A didactic connection to Kalman filtering theory is provided for a methodological overview.

1. Introduction

The partial least squares regression (PLSR) algorithms of Wold and Martens provide alternative, powerful tools for handling ill-conditioned multivariate regression data, see e.g. Martens and Næs (1989) and Höskuldson (1996) for overviews of the fundamental multivariate calibration concept, in which are presented the above two slightly different versions of the PLSR method. Both algorithms assume collinear regressor variables generated by underlying latent variables, and although they are different with respect to the score and loading matrices involved, they give identical predictors.

There is certainly no practical need for yet another algorithm that provide the same predictor as the well-known algorithms of Wold and Martens. However, these algorithms, as presented in the chemometrical literature, make no use of the dynamic systems theory available for readers from the control community, and the main aim of the present paper is therefore to provide a predictor derivation at first especially for this category of readers.

This actually also results in a simplified version of the Martens prediction algorithm making use of the so-called loading weight vectors only, which would appear to be of general interest (see Esbensen (2000) for a discussion of loadings and loading weights). The score vectors, which are also essential elements of chemometrics, are easily computed once a preliminary predictor is found.

The new algorithm has a lot in common with an algorithm developed by Helland (1988), and a similar algorithm by Di Ruscio (2000). Again, the reason behind the present modifications is mainly didactic. The key step in the simplification is the use of an explicit latent variables model, which facilitates an early introduction of the Helland predictor form using only loading weight vectors. The present exposition, we believe, constitute a novel, easy, and complete introduction to the prediction aspect of multivariate calibration.

2. The multivariate calibration problem

Assuming a standard regression problem with a scalar response variable y and multivariate regressor variables $\mathbf{x}$, the PLS1 setting in chemometrics, the object is to find a predictor $\hat{\mathbf{b}}$ from empirical or experimental data, that may be used to predict a new response y_0 from new observations $\mathbf{x}_0$ according to

$$\hat{y}_0 = \mathbf{x}_0^T \hat{\mathbf{b}} \tag{1}$$

The specific multivariate calibration problem arises when the number of $\mathbf{x}$ variables is larger than the number of observations in the available data, which calls for regularized solutions.

Remark 1 *In typical chemometric applications the regressor variables represent more or less noisy measurements. From a strict systems engineering point of view the notation y_1 for the response and $\mathbf{y}_2$ for the regressors would then be more natural. However, here we shall use the well-established chemometrical notation.*

Martens and Næs (1989) give an abundance of didactic multivariate calibration problems as seen from the chemometric point-of-view when dealing with what might be called static multivariate calibration (see Ergon (1998, 1999) for discussions of dynamic counterparts). One illustrative archetypal example of modern analytical instrument multivariate calibration is that of prediction of protein content in whole wheat kernels based on near infrared (NIR) spectroscopy (Norris, 1993). Here, the protein content is the response variable y, while the instrumental NIR reflectance at a large number of frequencies serve as the $\mathbf{x}$ variables. This example actually represents a modern breakthrough for practical chemometric multivariate calibration, allowing rapid NIR to replace the traditional slow, wet-chemical determinations. The fundamental problem in such cases is that the number of $\mathbf{x}$ variables may be much larger than the actual number of observations in the calibration (training) data, which gives rise to well-known statistical degrees-of-freedom problems. Also, the $\mathbf{X}$ data matrix typically is made up of significantly collinear variables, each of which is usually also fraught with a non-trivial measurement error. This case clearly will spell disaster for e.g. a multivariate linear regression (MLR) approach.

3. Theory

Least squares estimation

Assuming experimental data from N observations, $\mathbf{y} = [y_1\ y_2 \ldots y_N]^T$ and $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2 \ldots \mathbf{x}_N]^T$, and independent observation errors, we find the least squares (LS) regression solution (e.g. Johnson and Wichern, 1992)

$$\hat{\mathbf{b}}_{\mathrm{LS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2}$$

With a large number p of $\mathbf{x}$ variables, this solution will be very noise and collinearity sensitive, and in practical applications the LS method will give satisfactory results only when p is well below N. A general and detailed analysis of this problem for limited values of N is beyond the scope of the present paper (see e.g. Belsley (1991), but the issue is also well dealt with in the chemometrical literature).

Latent variables model

When, as indeed in very many practical multivariate calibration situations, the large number of $\mathbf{x}$ variables are significantly to highly collinear, the regressor information may be compressed into a much smaller number of *latent variables*

$\boldsymbol{\tau} = [\tau_1 \ \tau_2 \ldots \tau_a]^T$ (e.g. Burnham *et al.* (1999), Martens and Næs (1989) and Høskuldson (1996)). The model underlying such data compression is

$$\begin{aligned} \boldsymbol{\tau}_{k+1} &= \mathbf{e}_k \\ y_k &= \mathbf{C}_1 \boldsymbol{\tau}_k + v_{1,k} \\ \mathbf{x}_k &= \mathbf{W} \boldsymbol{\tau}_k + \mathbf{v}_{2,k} \end{aligned} \tag{3}$$

where $\mathbf{e}_k$, $v_{1,k}$ and $\mathbf{v}_{2,k}$ are white noise sequences with covariances $\mathbf{R}_e$, r_{11} and $\mathbf{R}_{22}$, and where we assume centered data. This is a special case of a more general dynamic model with $\boldsymbol{\tau}_{k+1} = \mathbf{A}\boldsymbol{\tau}_k + \mathbf{e}_k$, i.e. we use here $\mathbf{A} = \mathbf{0}$. We will also assume that $\mathbf{W}$ is orthonormal, i.e. that $\mathbf{W}^T\mathbf{W} = \mathbf{I}$.

With N observations and $\mathbf{T} = [\boldsymbol{\tau}_1 \ \boldsymbol{\tau}_2 \ldots \boldsymbol{\tau}_N]^T$, $\mathbf{v}_1 = [v_{11} \ v_{12} \ldots v_{1N}]^T$ and $\mathbf{V}_2 = [\mathbf{v}_{21} \ \mathbf{v}_{22} \ldots \mathbf{v}_{2N}]^T$, the latent variables model (3) gives the output equations

$$\mathbf{y} = \mathbf{T}\mathbf{C}_1^T + \mathbf{v}_1 \tag{4}$$

$$\mathbf{X} = \mathbf{T}\mathbf{W}^T + \mathbf{V}_2 \tag{5}$$

Remark 2 *If the regressor data are actually generated from an underlying state vector* $\mathbf{z}_k$ *as* $\mathbf{x}_k = \mathbf{C}_2\mathbf{z}_k + \mathbf{v}_{2,k}$, *the output equation (5) will be replaced by* $\mathbf{X} = \mathbf{Z}\mathbf{C}_2^T + \mathbf{V}_2$. *In the special noise free case with* $\mathbf{V}_2 = \mathbf{0}$, *and with* $N \to \infty$, *we may then find* $\mathbf{T}$ *and* $\mathbf{W}^T$ *by a similarity transformation based on factorization of* $\mathbf{X}$ *by a number of alternative methods, e.g. PCA and PLSR as described below. In practice we will always observe some noise and have a limited N, and factorization of* $\mathbf{X}$ *then gives only estimates* $\hat{\mathbf{T}}$ *and* $\hat{\mathbf{W}}^T$.

Regularized solutions

The LS solution of (5) is

$$\hat{\mathbf{T}} = \mathbf{X}\mathbf{W} \tag{6}$$

and from equations (4) and (6) we thus find the LS predictor related to the latent variables

$$\hat{\mathbf{C}}_1^T = (\hat{\mathbf{T}}^T\hat{\mathbf{T}})^{-1}\hat{\mathbf{T}}^T\mathbf{y} = (\mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{X}^T\mathbf{y} \tag{7}$$

which results in fitted experimental responses according to equation (4)

$$\hat{\mathbf{y}} = \hat{\mathbf{T}}\hat{\mathbf{C}}_1^T = \mathbf{X}\mathbf{W}\hat{\mathbf{C}}_1^T = \mathbf{X}\mathbf{W}(\mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{X}^T\mathbf{y} \tag{8}$$

and predictions of new responses

$$y_0 = \hat{\boldsymbol{\tau}}_0^T\hat{\mathbf{C}}_1^T = \mathbf{x}_0^T\mathbf{W}\hat{\mathbf{C}}_1^T = \mathbf{x}_0^T\mathbf{W}(\mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{X}^T\mathbf{y} \tag{9}$$

The regularized latent variables predictor thus becomes

$$\hat{\mathbf{b}}_{\mathrm{LV}} = \mathbf{W}(\mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{X}^T\mathbf{y} \tag{10}$$

This predictor is also given in Helland (1988), although not directly based on an LV model.

The problem now is to find $\mathbf{W}$ or more realistically good estimates $\hat{\mathbf{W}}$, and in this endeavour we have in fact a number of possibilities. A simple choice is $\hat{\mathbf{W}} = \mathbf{I}_p$, which brings us back to the LS solution (2). Other choices give the standard statistical PCR and standard chemometrical PLSR solutions as discussed below. The interested reader might note that the theoretically optimal regularization is given by a Kalman gain (see Appendix A for details).

Principal component regression

In PCR the weighting matrix estimate is $\hat{\mathbf{W}} = \hat{\mathbf{W}}_{\mathrm{PCR}} = \hat{\mathbf{P}}$, where $\hat{\mathbf{P}}$ is the loading matrix related to a principal components decomposition of $\mathbf{X}$ (Johnson and Wichern, 1989). We may find $\hat{\mathbf{T}} = \hat{\mathbf{T}}_{\mathrm{PCR}} = \mathbf{U}_1\mathbf{S}_1$ and $\hat{\mathbf{P}} = \mathbf{V}_1$ from the singular value decomposition

$$\begin{aligned}\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T = [\mathbf{U}_1 \;\; \mathbf{U}_2]\begin{bmatrix}\mathbf{S}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{S}_2\end{bmatrix}\begin{bmatrix}\mathbf{V}_1^T\\ \mathbf{V}_2^T\end{bmatrix} \\ = \mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^T + \mathbf{U}_2\mathbf{S}_2\mathbf{V}_2^T = \hat{\mathbf{T}}_{\mathrm{PCR}}\hat{\mathbf{P}}^T + \mathbf{E}\end{aligned} \tag{11}$$

The latent variables represented by the score matrix $\hat{\mathbf{T}}_{\mathrm{PCR}}$ are thus based on only $\mathbf{X}$ information.

In chemometrics there is a strong tradition for using the NIPALS algorithm for this decomposition (e.g. Martens and Næs, 1989). In the expression

$$\mathbf{X} = \hat{\mathbf{T}}_{\mathrm{PCR}}\hat{\mathbf{P}}^T + \mathbf{E} = \hat{\mathbf{t}}_1\hat{\mathbf{p}}_1^T + \hat{\mathbf{t}}_2\hat{\mathbf{p}}_2^T + \ldots + \hat{\mathbf{t}}_a\hat{\mathbf{p}}_a^T + \mathbf{E}_a \tag{12}$$

we then successively maximize the sample variances $\hat{\mathbf{t}}_1^T\hat{\mathbf{t}}_1/(N-1)$, $\hat{\mathbf{t}}_2^T\hat{\mathbf{t}}_2/(N-1)$ under the constraint that $\hat{\mathbf{t}}_2$ is orthogonal to $\hat{\mathbf{t}}_1$, $\hat{\mathbf{t}}_3^T\hat{\mathbf{t}}_3/(N-1)$ under the constraints that $\hat{\mathbf{t}}_3$ is orthogonal to $\hat{\mathbf{t}}_1$ and $\hat{\mathbf{t}}_2$ etc.

Partial least squares regression

Some of the latent variables represented in $\hat{\mathbf{T}}_{\mathrm{PCR}} = [\hat{\mathbf{t}}_1 \;\; \hat{\mathbf{t}}_2 \ldots \hat{\mathbf{t}}_a]$ may be weakly correlated with the response variable in $\mathbf{y}$. This is where the PLSR solution suggests to use *both* $\mathbf{X}$ and $\mathbf{y}$ information in order to find an improved version of $\hat{\mathbf{W}} = [\hat{\mathbf{w}}_1 \;\; \hat{\mathbf{w}}_2 \ldots \hat{\mathbf{w}}_a]$, the so-called *loading weight* matrix. In the Wold and Martens algorithms this is done by a step-wise computation of $\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_a$ (e.g. Martens and Næs, 1989), but a one-step procedure is also available (Di Ruscio, 2000).

The so-called Martens algorithm is based on the factorization

$$\mathbf{X} = \hat{\mathbf{T}}_{\mathrm{PLS}}\hat{\mathbf{W}}_{\mathrm{PLS}}^T + \mathbf{E} = \hat{\mathbf{t}}_1\hat{\mathbf{w}}_1^T + \hat{\mathbf{t}}_2\hat{\mathbf{w}}_2^T + \ldots + \hat{\mathbf{t}}_a\hat{\mathbf{w}}_a^T + \mathbf{E}_a \tag{13}$$

and the following modified algorithm has the same starting point.

Remark 3 *The Martens algorithm uses a non-orthogonal score matrix, i.e.* $\hat{\mathbf{T}}_{\mathrm{PLS}}^T\hat{\mathbf{T}}_{\mathrm{PLS}}$ *is non-diagonal, while the* alternative *Wold algorithm makes use of an orthogonal score matrix (e.g. Martens and Næs, 1989).*

In the first step we try to explain $\mathbf{y}$ by use of only one component in equation (13), i.e. by using

$$\hat{\mathbf{X}} = \hat{\mathbf{t}}_1\hat{\mathbf{w}}_1^T + \mathbf{E}_1 \tag{14}$$

where it follows from equation (6) that

$$\hat{\mathbf{t}}_1 = \mathbf{X}\hat{\mathbf{w}}_1 \tag{15}$$

We then maximize the sample covariance $\hat{\mathbf{t}}_1^T\mathbf{y}/(N-1) = \hat{\mathbf{w}}_1^T\mathbf{X}_y^T/(N-1)$ under the constraint that $\hat{\mathbf{w}}_1$ has the length $\sqrt{\hat{\mathbf{w}}_1^T\hat{\mathbf{w}}_1} = 1$ (instead of maximizing $\hat{\mathbf{t}}_1^T\hat{\mathbf{t}}_1/(N-1)$ as in PCR). We find the maximum when $\hat{\mathbf{w}}_1$ has the same direction as $\mathbf{X}^T\mathbf{y}$, i.e.

$$\hat{\mathbf{w}}_1 = c_1\mathbf{X}^T\mathbf{y} \tag{16}$$

where $c_1 = (\mathbf{y}^T\mathbf{X}\mathbf{X}^T\mathbf{y})^{-1/2}$.

Remark 4 *The scaling of* $\hat{\mathbf{w}}_1$ *is not absolutely necessary, i.e. we may use* $c_1 = 1$ *(Helland, 1988), but it is very often carried out for practical reasons, and it furthers the most comprehensive interpretation possibilities according to the chemometric tradition.*

Remark 5 *Maximization of the sample covariance is the favored* ad hoc *chemometric solution. The theoretically optimal solution assuming known covariances* $\mathbf{R}_e$ *and* $\mathbf{R}_{22}$ *is to let* $\mathbf{W}$ *be a transposed Kalman gain (see Appendix A).*

The result of this first step is $\hat{\mathbf{W}}_{\text{PLS}} = \hat{\mathbf{W}}_1 = \hat{\mathbf{w}}_1$, and the fitted responses according to equation (8)

$$\hat{\mathbf{y}}_1 = \mathbf{X}\hat{\mathbf{W}}_1(\hat{\mathbf{W}}_1^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{W}}_1)^{-1}\hat{\mathbf{W}}_1^T\mathbf{X}^T\mathbf{y} \tag{17}$$

We also obtain the residual (this step is often called "updating" or "deflation")

$$\boldsymbol{\varepsilon}_1 = \mathbf{y} - \hat{\mathbf{y}}_1 = \mathbf{y} - \mathbf{X}\hat{\mathbf{W}}_1(\hat{\mathbf{W}}_1^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{W}}_1)^{-1}\hat{\mathbf{W}}_1^T\mathbf{X}^T\mathbf{y} \tag{18}$$

What has just been completed for the first PLSR component can now be iterated, resulting in the next, orthogonal component. Thus in the second step we try to explain the residual $\boldsymbol{\varepsilon}_1$ by using the second component in equation (13), i.e. by maximizing the sample covariance $\hat{\mathbf{t}}_2^T\boldsymbol{\varepsilon}_1/(N-1) = \hat{\mathbf{w}}_2^T\mathbf{X}^T\boldsymbol{\varepsilon}_1/(N-1)$ under the constraints $\sqrt{\hat{\mathbf{w}}_2^T\hat{\mathbf{w}}_2} = 1$ and $\hat{\mathbf{w}}_2^T\hat{\mathbf{w}}_1 = 0$. The result now is

$$\hat{\mathbf{w}}_2 = c_2\mathbf{X}^T\boldsymbol{\varepsilon}_1 \tag{19}$$

where $c_2 = (\boldsymbol{\varepsilon}_1^T\mathbf{X}\mathbf{X}^T\boldsymbol{\varepsilon}_1)^{-1/2}$. Here is remains to prove that $\hat{\mathbf{w}}_2^T\hat{\mathbf{w}}_1 = 0$ (taken care of in Appendix B).

After this second step we have $\hat{\mathbf{W}}_{\text{PLS}} = \hat{\mathbf{W}}_2 = [\hat{\mathbf{w}}_1 \ \hat{\mathbf{w}}_2]$, the fitted responses according to equation (8)

$$\hat{\mathbf{y}}_2 = \mathbf{X}\hat{\mathbf{W}}_2(\hat{\mathbf{W}}_2^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{W}}_2)^{-1}\hat{\mathbf{W}}_2^T\mathbf{X}^T\mathbf{y} \tag{20}$$

and the residual

$$\boldsymbol{\varepsilon}_2 = \mathbf{y} - \hat{\mathbf{y}}_2 \tag{21}$$

This contemporary residual, $\boldsymbol{\varepsilon}_2$, is used as input to the third step etc. We thus use more and more detailed factorizations of $\mathbf{X}$ (adding new PLSR components) in order to explain consecutive residuals of $\mathbf{y}$.

4. A new didactic prediction algorithm

The algorithm developed above is as follows (where the scaling in equation (23) may be omitted):

1. Set $a = 1$ and $\varepsilon_0 = \mathbf{y}$.
2. Compute

$$\hat{\mathbf{w}}_a = \mathbf{X}^T \varepsilon_{a-1} \tag{22}$$

$$\hat{\mathbf{w}}_a \leftarrow \frac{\hat{\mathbf{w}}_a}{\sqrt{\hat{\mathbf{w}}_a^T \hat{\mathbf{w}}_a}} \tag{23}$$

$$\hat{\mathbf{W}}_a = [\hat{\mathbf{w}}_1 \ \hat{\mathbf{w}}_2 \ldots \hat{\mathbf{w}}_a] \tag{24}$$

$$\hat{\mathbf{b}}_a = \hat{\mathbf{W}}_a(\hat{\mathbf{W}}_a^T \mathbf{X}^T \mathbf{X} \hat{\mathbf{W}}_a)^{-1} \hat{\mathbf{W}}_a^T \mathbf{X}^T \mathbf{y} \tag{25}$$

   and

$$\varepsilon_a = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}}_a \tag{26}$$

3. Let $a \leftarrow a + 1$ and go to step 2.

The algorithm thus produces a sequence of predictors $\hat{\mathbf{b}}_1$, $\hat{\mathbf{b}}_2$ etc. The appropriate number of column vectors $\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_A$ to use is of critical importance. In general it would be unavoidable to either underfit or overfit this modeling without an unambiguous stopping rule, i.e. the “optimal number of PLSR components” must be decided upon by a suitable evaluation of “the modeling fit”. As Höskuldson (1996) has pointed out, this optimization criterion must include terms which reflect both the **X**-modeling fit as well as prediction error minimization—in fact he devised a new compound principle, the H-principle, for obtaining the optimal balance between these two terms, *ibid.*

In the chemometric practise, there has been developed a tradition for using an empirical test for finding the “optimal prediction strength” via a suitable validation procedure, which is often of the cross-validation form, e.g. Martens and Næs (1989). But Esbensen (2000) and Esbensen and Huang (2001) have been adamant in pointing out many deficiencies regarding cross-validation, while demonstrating the almost universal need for validation against an independent, so-called “test data set” (the test set validation imperative). A proper validation is an essential part of any multivariate data modeling, not just prediction modeling, *ibid.*

Remark 6 *The new algorithm utilizes residuals of* **y** *in order to find consecutive loading weight vectors* $\hat{\mathbf{w}}_a$. *The Wold and Martens algorithms use residuals of* **X** *as well, and this may also be included here. However, this may seen somewhat contrived and confusing when the goal is to explain* **y** *in more and more detail by use of the information available in* **X**, *and in the present algorithm it is in any case unnecessary.*

A very significant part of practical chemometrics is interpretation based on graphical modeling, *viz.* the score and loading weight matrices etc., Esbensen (2000). These model results must also be computed, either as a part of the algorithm (for the entire series of components calculated), or separately, i.e. for specific values of a

(the optimal number of PLS-components). We can find the non-orthogonal score matrix for $\mathbf{X}$ as

$$\hat{\mathbf{T}}_a = \mathbf{X}\hat{\mathbf{W}}_a \tag{27}$$

and the loading vector for $\mathbf{y}$ (see Martens and Næs (1989)) as

$$\hat{\mathbf{q}}_a = (\hat{\mathbf{W}}_a^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{W}}_a)^{-1}\hat{\mathbf{W}}_a^T\mathbf{X}^T\mathbf{y} \tag{28}$$

5. Conclusion

A combined version of the Martens and Helland PLSR algorithms based on an explicit latent variables model and making use of only loading weight vectors $\hat{\mathbf{w}}_a$ is developed from a novel didactic perspective. The resulting predictor is identical to the Helland predictor, and it is furthermore equivalent with the Wold and Martens predictors. Estimates of the non-orthogonal score matrix $\hat{\mathbf{T}}$ and the loading vector $\hat{\mathbf{q}}$ may also be computed as desired.

It is emphasized that we are here exclusively dealing with the so-called non-orthogonal PLSR algorithms, which does not allow a similar insight into the specific PLSR models as can be achieved by using the alternative so-called Wold-models, in which $\hat{\mathbf{T}}$ is indeed orthogonal, as is $\hat{\mathbf{W}}$ (Appendix B). It has been pointed out in chemometrics, that exclusion of this latter information may at times severely cripple the usefulness of the PLSR models with respect to all other aspects than mere prediction, Esbensen (2000). While this may not appear to be of specific importance to the control community, it certainly is in the static chemometric realm. We shall address these aspects in a companion paper.

A connection to Kalman filtering theory is given in Appendix A.

Appendix A

Basis in Kalman filtering theory

The general static case

In the latent variables model (3) the transition matrix is $\mathbf{A} = 0$, and there is no manipulated input $\mathbf{u}_k$. The Kalman filter prediction (a priori) state estimate is therefore $\hat{\boldsymbol{\tau}}_{k|k-1} = \mathbf{0}$ and thus $E(\boldsymbol{\tau}_k - \hat{\boldsymbol{\tau}}_{k|k-1})\ (\boldsymbol{\tau}_k - \hat{\boldsymbol{\tau}}_{k|k-1})^T = E\boldsymbol{\tau}_k\boldsymbol{\tau}_k^T = \mathbf{R}_e$. Thus, in a Kalman filter driven by $\mathbf{x}_k$ the updated (a posteriori) state estimate is

$$\hat{\boldsymbol{\tau}}_{k|k} = \mathbf{K}_2\mathbf{x}_k \tag{29}$$

where

$$\mathbf{K}_2 = \mathbf{R}_e\mathbf{W}^T(\mathbf{W}\mathbf{R}_e\mathbf{W}^T + \mathbf{R}_{22})^{-1} \tag{30}$$

(e.g. Grewal and Andrews, 1993). This intermediate result was first presented in Berntsen (1988).

The resulting response estimate is

$$\hat{y}_k = \mathbf{C}_1\mathbf{K}_2\mathbf{x}_k \tag{31}$$

i.e.

$$y_k = \mathbf{C}_1\mathbf{K}_2\mathbf{x}_k + \eta_k \tag{32}$$

where it can be shown that η_k is white noise (Ergon 1998, 1999). Assuming $\mathbf{K}_2$ known,

$\mathbf{C}_1$ unknown and experimental data available, we may from this find the LS estimate corresponding to equation (7)

$$\hat{\mathbf{C}}_1^T = (\mathbf{K}_2\mathbf{X}^T\mathbf{X}\mathbf{K}_2^T)^{-1}\mathbf{K}_2\mathbf{X}^T\mathbf{y} \tag{33}$$

and thus the fitted primary outputs corresponding to equation (8)

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{K}_2^T\hat{\mathbf{C}}_1^T = \mathbf{X}\mathbf{K}_2^T(\mathbf{K}_2\mathbf{X}^T\mathbf{X}\mathbf{K}_2^T)^{-1}\mathbf{K}_2\mathbf{X}^T\mathbf{y} \tag{34}$$

The optimal choice of $\mathbf{W}$ in the Helland predictor equation (10) is thus in theory $\mathbf{W} = \mathbf{K}_2^T$. This connection between the regularized least squares solution and Kalman filtering appears to be a parallel to the connection between a regularized solution of a convolution integral and Wiener filtering presented by Tikhonov and Arsenin (1977). However, in practice we must be content with e.g. $\mathbf{W} = \hat{\mathbf{W}}_{\mathrm{PLS}}$.

Special noise free case

An estimate of $\mathbf{R}_e$ is

$$\hat{\mathbf{R}}_e = \frac{1}{\mathrm{N}-1}\hat{\mathbf{T}}^T\hat{\mathbf{T}} \tag{35}$$

which may be inserted in equation (30). With $\mathbf{R}_{22} = 0$ and after multiplication with $\hat{\mathbf{W}}^T\hat{\mathbf{W}} = \mathbf{I}$ we then obtain

$$\hat{\mathbf{K}}_2 = \hat{\mathbf{W}}^T\hat{\mathbf{W}}\hat{\mathbf{T}}^T\hat{\mathbf{T}}\hat{\mathbf{W}}^T(\hat{\mathbf{W}}\hat{\mathbf{T}}^T\hat{\mathbf{T}}\hat{\mathbf{W}}^T)^{-1} = \hat{\mathbf{W}}^T \tag{36}$$

In this special case the loading weight matrix $\hat{\mathbf{W}}$ thus represents the best approximation of the Kalman gain that can be obtained from from the available data. In the general case an improved estimate $\hat{\mathbf{K}}_2$ would require some information concerning the covariance matrix $\mathbf{R}_{22}$ (Ergon and Esbensen, 2001).

Appendix B

Proof of orthogonality of weighting vectors

It is necessary to prove that $\hat{\mathbf{w}}_{a+1}$ is orthogonal to $\hat{\mathbf{w}}_a$, $\hat{\mathbf{w}}_{a-1}$, ..., $\hat{\mathbf{w}}_1$, i.e. that $\hat{\mathbf{W}}_a^T\hat{\mathbf{w}}_{a+1} = \mathbf{0}$. From equations (22) to (26) follows

$$\begin{aligned}
\hat{\mathbf{W}}_a^T\hat{\mathbf{w}}_{a+1} &= \hat{\mathbf{W}}_a^T c_{a+1}\mathbf{X}^T\boldsymbol{\varepsilon}_a \\
&= \hat{\mathbf{W}}_a^T\mathbf{X}^T c_{a+1}[\mathbf{y} - \mathbf{X}\hat{\mathbf{W}}_a(\hat{\mathbf{W}}_a^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{W}}_a)^{-1}\hat{\mathbf{W}}_a^T\mathbf{X}^T\mathbf{y}] \\
&= c_{a+1}\hat{\mathbf{W}}_a^T\mathbf{X}^T\mathbf{y} - c_{a+1}\hat{\mathbf{W}}_a^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{W}}_a(\hat{\mathbf{W}}_a^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{W}}_a)^{-1}\hat{\mathbf{W}}_a^T\mathbf{X}^T\mathbf{y} \\
&= c_{a+1}\hat{\mathbf{W}}_a^T\mathbf{X}^T\mathbf{y} - c_{a+1}\hat{\mathbf{W}}_a^T\mathbf{X}^T\mathbf{y} = 0
\end{aligned} \tag{37}$$

References

Belsley, D. A. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression*, Wiley, New York.

Berntsen, H. (1988). *Utvidet Kalmanfilter og multivariabel kalibrering*, Report STF48 A88019, SINTEF, Trondheim, Norway.

Di Ruscio, D. (2000). *A. weighted view on the partial least squares algorithm*, Automatica 36, pp. 831–850.

ERGON, R. (1998). *Dynamic system multivariate calibration by system identification methods*, Modeling, Identification and Control, Vol. 19, No. 2, pp. 77–97.

ERGON, R. (1999). *Dynamic System Multivariate Calibration for Optimal Primary Output Estimation*, Ph.D. thesis, The Norwegian University of Science and Technology and Telemark University College, Norway.

ERGON, R. and ESBENSEN, K. H. (2001). *Static PLSR optimization based on Kalman filtering theory and noise covariance estimation*, 7th Scandinavian Symposium on Chemometrics, Copenhagen.

ESBENSEN, K. (2000). *Multivariate Data Analysis – in practice*, Camo ASA, Trondheim, Norway.

GREWAL, M. S. and ANDREWS, A. P. (1993). *Kalman Filtering: Theory and Practice*, Prentice Hall, New Jersey.

ESBENSEN, K. and HUANG, J. (2001). *Principles of proper validation*, in preparation for Journal of Chemometrics.

HELLAND, I. S. (1988). On the structure of partial least squares regression, *Communications in Statistics*, 17(2), pp. 581–607.

HØSKULDSSON, A. (1996). *Prediction Methods in Science and Technology*, Thor Publishing, Copenhagen.

JOHNSON, R. A. and WICHERN, D. W. (1992). *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey.

MARTENS, H. and NÆS, T. (1989). *Multivariate Calibration*, John Wiley & Sons, New York.

NORRIS, K. H. (1993). Extracting information from spectrophotometric curves. Predicting chemical composition from visible and near-infrared spectra, *Proc. IUFost Symp. Food Research and Data Analysis*, Sept. 1982, Oslo, Norway (MARTENS and RUSSWORM, eds.), Applied Science Publ., 95–113.

TICHONOV, A. N. and ARSENIN, V. Y. (1977). *Solutions of Ill-Posed Problems*, V. H. Winston & Sons, Washington, DC.