Constrained numerical optimization of PCR/PLSR predictors

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

Published in Chemometrics and Intelligent Laboratory Systems 65 (2003) 293-303

Abstract

Assuming a fully known latent variables model, the optimal multivariate calibration predictor is found from Kalman filtering theory. From this follows the best possible column space for a loading weight matrix $\mathbf{W}_{\text{opt.}}$ in a predictor based on the latent variables, and thus the optimal factorization of the regressor matrix $\mathbf{X}$. Although the optimal predictor cannot be directly determined in a practical case, we may still make an attempt to find it. The paper presents a simple algorithm for a constrained numerical search for a $\mathbf{W}_{\text{opt.}}$ matrix spanning the optimal column space, using a principal component analysis (PCR) or a partial least squares (PLS) factorization as a starting point. The constraint is necessary in order to avoid overfitting, and it is based on an assumption of a smooth predictor. A simulation example and data from a metal ion mixture experiment are used to demonstrate the feasibility of the proposed method.

1 Introduction

The aim of this paper is to show that multivariate calibration results from principal component regression (PCR) or partial least squares regression (PLSR) at least in some cases may be improved by a numerical search for an optimal factorization of the $\mathbf{X}$ data matrix, i.e. for an optimal loading or loading weight matrix. The theoretical basis for this is found in general Kalman filtering theory, and the fact that an optimal factorization under the assumptions of a linear latent variables (LV) model and normal LV and $\mathbf{X}$-noise distributions can be shown to exist. An additional assumption is that the resulting optimal predictor is smooth, such that a numerical search for an optimal factorization can be constrained by use of a predictor roughness index. This is necessary in order to avoid convergence to the least squares solution and thus overfitting. The treatment is limited to the scalar response case.

For an introduction, assume an underlying LV model

$$y_k = \mathbf{Q}\boldsymbol{\tau}_k + f_k \tag{1}$$
$$\mathbf{x}_k = \mathbf{L}\boldsymbol{\tau}_k + \mathbf{e}_k, \tag{2}$$

with a scalar response variable y_k, regressor variables $\mathbf{x}_k = \mathbb{R}^{p\times 1}$, latent variables $\boldsymbol{\tau}_k = \mathbb{R}^{A\times 1}$, $\mathbf{Q} \in \mathbb{R}^{1\times A}$ and $\mathbf{L} = \mathbb{R}^{p\times A}$, where $\mathbf{L}$ has orthonormal columns. The error terms f_k and $\mathbf{e}_k$ are assumed to be independent with expected variance $r_f = E\left\{f_k^2\right\}$ and covariance $\mathbf{R}_e = E\left\{\mathbf{e}_k\mathbf{e}_k^T\right\}$, while $\boldsymbol{\tau}_k$ has the expected covariance $\mathbf{R}_\tau = E\left\{\boldsymbol{\tau}_k\boldsymbol{\tau}_k^T\right\}$. With data collected from $k = 1, 2, \ldots, N$ observations we from

this obtain the sample latent variables model [1]

$$\mathbf{y} = \mathbf{T}\mathbf{Q}^T + \mathbf{f} \tag{3}$$
$$\mathbf{X} = \mathbf{T}\mathbf{L}^T + \mathbf{E}. \tag{4}$$

Assuming $A << p$ the multivariate calibration predictor based on a given modeling set can then be expressed as [2]

$$\hat{\mathbf{b}}_{\mathrm{LV}} = \hat{\mathbf{L}}\left(\hat{\mathbf{L}}^T\mathbf{X}^T\mathbf{X}\hat{\mathbf{L}}\right)^{-1}\hat{\mathbf{L}}^T\mathbf{X}^T\mathbf{Y}. \tag{5}$$

This applies to both PCR and PLSR, with $\hat{\mathbf{L}}$ being the loading matrix $\hat{\mathbf{P}}$ or the loading weight matrix $\hat{\mathbf{W}}$ respectively [3].

It has further been shown that the theoretically optimal $\hat{\mathbf{L}}$ matrix is a transposed Kalman gain $\mathbf{K}^T$ [4], or any matrix spanning the same column space. This matrix can be computed only if both $\mathbf{R}_\tau$ and $\mathbf{R}_e$ are known, which of course they are not in practical cases. However, it has also been shown that an estimate $\hat{\mathbf{K}}$ may be found by means of covariance estimation using extra $\mathbf{X}$-observations, i.e. from a long $\mathbf{X}$ matrix [4]. When this is applied to PCR it turns out that $\hat{\mathbf{L}}$ in (5) is replaced by the loading matrix $\hat{\mathbf{P}}_{\mathrm{long}}$ found from $\mathbf{X}_{\mathrm{long}}$, which method for stabilization of the PCR predictor was earlier presented in [5]. Application to PLSR may, however, give better prediction results and/or fewer components.

A singular value decomposition (SVD) (or any other method for obtaining an orthonormal basis, e.g. QR decomposition) of the transposed Kalman gain gives

$$\mathbf{K}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix}\begin{bmatrix} \mathbf{S}_1 \\ \mathbf{0} \end{bmatrix}\mathbf{V}^T = \mathbf{U}_1\mathbf{S}_1\mathbf{V}^T, \tag{6}$$

where $\mathbf{S}_1\mathbf{V}^T \in \mathbb{R}^{A\times A}$ is a square and invertible matrix. The matrix $\hat{\mathbf{L}}$ in (5) may thus be replaced by $\mathbf{U}_1 \in \mathbb{R}^{p\times A}$, which just as $\hat{\mathbf{P}}$ (PCR) and $\hat{\mathbf{W}}$ (PLSR) has orthonormal columns. It may in fact be replaced by any orthonormal matrix $\mathbf{W}_{\mathrm{opt.}}$ with the same column space as $\mathbf{U}_1$, i.e. an optimization of the predictor may be performed by an column space adaptation using e.g. $\hat{\mathbf{W}}$ as a starting point.

The present paper investigates the possibilities to find a $\mathbf{W}_{\mathrm{opt.}}$ matrix by a numerical search using cautiously modified $\hat{\mathbf{W}}$ (or $\hat{\mathbf{P}}$) matrices, i.e. without using extra $\mathbf{X}$ observations as in [4]. When this is attempted, two principal problems are encountered:

- Optimization using only calibration/modeling data, i.e. finding $\min\{\mathrm{RMSEC}\}$, will result in overfitting and poor prediction results for new $\mathbf{X}$ data.
- Optimization using also validation data, i.e. finding $\min\{\mathrm{RMSEP}\}$ for a given test set, makes the test set a part of the modeling set, again with overfitting as result. The corresponding overfitting problem will occur also if cross-validation is used in the optimization algorithm.

These problems can be overcome only by use of some optimization constraints. In this work we assume a theoretically smooth predictor, and use the roughness of (5) for this purpose, i.e. the search for $\mathbf{W}_{\mathrm{opt.}}$ is constrained by the requirement that a given predictor roughness index should not increase.

The theoretical basis presented above is somewhat expanded in Section 2, while a simple optimization algorithm including a roughness index is introduced in Section 3. A simulation example in Section 4 makes it possible to compare optimization results with a theoretically optimal solution. A real world data example involving metal ion mixtures is presented in Section 5, followed by a summary and conclusions in Section 6.

2 Theoretical basis

The Helland predictor

The PCR and PLSR regularizations are based on the latent variables model (3,4) above. The least squares (LS) solution of (4) is

$$\hat{\mathbf{T}} = \mathbf{X}\mathbf{L}\left(\mathbf{L}^T\mathbf{L}\right)^{-1}, \tag{7}$$

and from (3) and (7) we thus find the LS predictor related to the latent variables

$$\hat{\mathbf{Q}}^T = \left(\hat{\mathbf{T}}^T\hat{\mathbf{T}}\right)^{-1}\hat{\mathbf{T}}^T\mathbf{Y} = \left(\left(\mathbf{L}^T\mathbf{L}\right)^{-1}\mathbf{L}^T\mathbf{X}^T\mathbf{X}\mathbf{L}\left(\mathbf{L}^T\mathbf{L}\right)^{-1}\right)^{-1}\left(\mathbf{L}^T\mathbf{L}\right)^{-1}\mathbf{L}^T\mathbf{X}^T\mathbf{Y}, \tag{8}$$

which after some simplifications results in fitted experimental responses according to (3),

$$\hat{\mathbf{y}} = \hat{\mathbf{T}}\hat{\mathbf{Q}}^T = \mathbf{X}\mathbf{L}\left(\mathbf{L}^T\mathbf{X}^T\mathbf{X}\mathbf{L}\right)^{-1}\mathbf{L}^T\mathbf{X}^T\mathbf{Y}. \tag{9}$$

The regularized LV predictor thus becomes

$$\hat{\mathbf{b}}_{\mathrm{LV}} = \mathbf{L}\left(\mathbf{L}^T\mathbf{X}^T\mathbf{X}\mathbf{L}\right)^{-1}\mathbf{L}^T\mathbf{X}^T\mathbf{Y}. \tag{10}$$

This predictor was first presented in [2], although there not explicitly based on an LV model. Note that $\mathbf{L} = \hat{\mathbf{P}}$ gives the standard PCR predictor, while $\mathbf{L} = \hat{\mathbf{W}}$ gives the standard PLSR predictor [3].

The optimal predictor

In order to obtain a theoretical basis for both an optimization algorithm and simulation comparisons we need an optimal predictor formulation. The optimal predictor may be found by use of general Kalman filtering theory [6]. We will, however, derive the optimal solution directly by introduction of the optimal state estimate related to the LV model (1,2),

$$\hat{\boldsymbol{\tau}}_k = \mathbf{K}\mathbf{x}_k. \tag{11}$$

We chose $\mathbf{K}$ such that the expectation

$$\begin{aligned}\mathbf{R} &= E(\boldsymbol{\tau}_k - \hat{\boldsymbol{\tau}}_k)(\boldsymbol{\tau}_k - \hat{\boldsymbol{\tau}}_k)^T = E\left[\boldsymbol{\tau}_k - \mathbf{K}\left(\mathbf{L}\boldsymbol{\tau}_k + \mathbf{e}_k\right)\right]\left[\boldsymbol{\tau}_k - \mathbf{K}\left(\mathbf{L}\boldsymbol{\tau}_k + \mathbf{e}_k\right)\right]^T \\ &= (\mathbf{I} - \mathbf{K}\mathbf{L})E\left(\boldsymbol{\tau}_k\boldsymbol{\tau}_k^T\right)(\mathbf{I} - \mathbf{K}\mathbf{L})^T + \mathbf{K}E\left(\mathbf{e}_k\mathbf{e}_k^T\right)\mathbf{K}^T\end{aligned} \tag{12}$$

is minimized. Using $E\boldsymbol{\tau}_k\boldsymbol{\tau}_k^T = \mathbf{R}_\tau$ and $E\mathbf{e}_k\mathbf{e}_k^T = \mathbf{R}_e$ we find [7]

$$\frac{\partial}{\partial\mathbf{K}}trace(\mathbf{R}) = -2(\mathbf{I} - \mathbf{K}\mathbf{L})\mathbf{R}_\tau\mathbf{L}^T + 2\mathbf{K}\mathbf{R}_e, \tag{13}$$

i.e. $\frac{\partial}{\partial\mathbf{K}}trace(\mathbf{R}) = \mathbf{0}$ gives the optimal solution

$$\mathbf{K} = \mathbf{R}_\tau\mathbf{L}^T\left(\mathbf{L}\mathbf{R}_\tau\mathbf{L}^T + \mathbf{R}_e\right)^{-1}. \tag{14}$$

This intermediate result, derived from general Kalman filtering theory, was first presented in [8].

The resulting optimal response estimate is

$$\hat{\mathbf{y}}_k = \mathbf{Q}\mathbf{K}\mathbf{x}_k, \tag{15}$$

i.e. the optimal predictor is

$$\mathbf{b}_{\mathrm{KF}} = \mathbf{K}^T\mathbf{Q}^T = \left(\mathbf{L}\mathbf{R}_\tau\mathbf{L}^T + \mathbf{R}_e\right)^{-1}\mathbf{L}\mathbf{R}_\tau\mathbf{Q}^T. \tag{16}$$

Optimality here means that (16) gives the best linear unbiased estimate (BLUE), and the best possible estimate whatsoever assuming normal LV and $\mathbf{X}$-noise distributions [6]. This predictor will be used as a source of reference in the simulation example in Section 4. Note, however, that with noise that is not Gaussian, a biased and/or non-linear predictor may give even better results.

From (3) and (15) follows

$$\mathbf{y} = \mathbf{X}\hat{\mathbf{K}}^T\mathbf{Q}^T + \mathbf{f}_{\mathrm{KF}}, \tag{17}$$

where $\mathbf{f}_{\mathrm{KF}}$ is a random noise term, and assuming $\mathbf{Q}$ unknown an LS solution $\hat{\mathbf{Q}}^T$ thus results in

$$\hat{\mathbf{b}}_{\mathrm{KF}} = \mathbf{K}^T\hat{\mathbf{Q}}^T = \mathbf{K}^T\left(\mathbf{K}\mathbf{X}^T\mathbf{X}\mathbf{K}^T\right)^{-1}\mathbf{K}\mathbf{X}^T\mathbf{Y}. \tag{18}$$

This is the predictor (10) with $\mathbf{L}$ replaced by $\mathbf{K}^T$. As shown in (6) and the discussion that follows there, $\mathbf{K}^T$ may here be replaced by any orthonormal matrix $\mathbf{W}_{\mathrm{opt.}}$ spanning the same column space. Also this predictor will be used as a source of reference in the simulation example in Section 4.

Remark 1 *PCR and PLSR are often referred to as biased regression methods. Assuming the underlying latent variables model (3,4), this is meaningful only when comparing with the BLUE (16), or the LS counterpart (18).*

We summarize the theoretical basis in the following theorem, where we also include some score matrix computations.

Theorem 1

Assume the linear latent variables models (1,2) and (3,4) with $p >> A$, and the problem of predicting a new response from new observations according to $y_{\text{new}}^T = \mathbf{x}_{\text{new}}^T \hat{\mathbf{b}}$. The optimal predictor is then given by (16), which with $\mathbf{Q}$ unknown gives the optimal LS predictor

$$\hat{\mathbf{b}}_{\text{KF}} = \mathbf{W}_{\text{opt.}} \left(\mathbf{W}_{\text{opt.}}^T \mathbf{X}^T \mathbf{X} \mathbf{W}_{\text{opt.}} \right)^{-1} \mathbf{W}_{\text{opt.}}^T \mathbf{X}^T \mathbf{Y}, \tag{19}$$

where $\mathbf{W}_{\text{opt.}} \in \mathbb{R}^{p \times A}$ with orthonormal columns is found from the Kalman gain (14) by e.g. the SVD

$$\mathbf{K}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{0} \end{bmatrix} \mathbf{V}^T = \mathbf{U}_1 \mathbf{S}_1 \mathbf{V}^T = \mathbf{W}_{\text{opt.}} \mathbf{Z} \mathbf{Z}^{-1} \mathbf{S}_1 \mathbf{V}^T, \tag{20}$$

where $\mathbf{Z}$ is any suitable transformation matrix. The resulting predictor (19) is the best linear unbiased estimator (BLUE) in the LS sense (assuming $\mathbf{Q}$ unknown), and the best LS estimator whatsoever also assuming normal LV and $\mathbf{X}$-noise distributions.

An optimal factorization of $\mathbf{X}$ is

$$\mathbf{X} = \mathbf{T}_{\text{opt.}} \mathbf{W}_{\text{opt.}}^T + \mathbf{E}, \tag{21}$$

and an LS estimate of the non-orthogonal score matrix $\mathbf{T}_{\text{opt.}}$ [3] is given by

$$\hat{\mathbf{T}}_{\text{opt.}} = \mathbf{X} \mathbf{W}_{\text{opt.}}. \tag{22}$$

As shown in [9], a factorization with orthogonal score and loading weight matrices is furthermore given by the SVD

$$\hat{\mathbf{X}} = \hat{\mathbf{T}}_{\text{opt.}} \mathbf{W}_{\text{opt.}}^T = \begin{bmatrix} \hat{\mathbf{U}}_1 & \hat{\mathbf{U}}_2 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{S}}_1 \\ 0 \end{bmatrix} \hat{\mathbf{V}}^T \mathbf{W}_{\text{opt.}}^T = \left(\hat{\mathbf{U}}_1 \hat{\mathbf{S}}_1 \right) \left(\mathbf{W}_{\text{opt.}} \hat{\mathbf{V}} \right)^T = \hat{\mathbf{T}}_\perp \hat{\mathbf{W}}_\perp^T. \tag{23}$$

■

It follows from this theorem that under the assumptions given the optimal predictor (19) exists, and that the $\mathbf{W}_{\text{opt.}}$ matrix may be seen as a column space adapted version of the ordinary PLSR loading weight matrix $\hat{\mathbf{W}}$, or alternatively of the PCR loading matrix $\hat{\mathbf{P}}$.

3 Numerical random search algorithm

Modeling and validation RMSE values

The numerical search algorithm below involves the root mean square error of calibration, based on the available modeling data set and assuming a scalar response y_k,

$$\text{RMSEC} = \sqrt{\frac{1}{N} \sum_{k=1}^{N} \left(y_k - \hat{y}_k \right)^2}. \tag{24}$$

The corresponding root mean square error of prediction (RMSEP) based on an independent validation data set is used for the search interrupt decision only.

The idea behind the algorithm

The idea behind the algorithm is to search for an improved model fit by small random modifications of the PLSR loading weight matrix $\hat{\mathbf{W}}$ (or the PCR loading matrix $\hat{\mathbf{P}}$), and at the same time avoid overfitting by requiring that the initial smoothness of the predictor should not be impaired. The algorithm is based on a straightforward random search, while developement of a more effective algorithm is left for further work.

Theoretical limits

It follows from the theory above that the predictor (16) is the best predictor whatsoever assuming normal LV and **X**-noise distributions, and that (19) is then the best predictor that can be found through a numerical search for $\mathbf{W}_{\text{opt.}}$. For other distributions an even better biased and/or non-linear predictor can in theory be found.

The algorithm

1. Set $i = 0$ and use the available modeling data and an ordinary PLSR (or PCR) algorithm to find an initial loading weight matrix $\mathbf{W}_i = \hat{\mathbf{W}} \in \mathbb{R}^{p \times A}$ (or loading matrix $\mathbf{W}_i = \hat{\mathbf{P}} \in \mathbb{R}^{p \times A}$). Find the corresponding predictor $\hat{\mathbf{b}}_i$ according to (5) with $\hat{\mathbf{L}} = \mathbf{W}_i$, and the RMSEC_i value according to (24). Also compute the RMSEP_i value based on an independent validation data set (to be used for iteration interrupt only).

2. Compute a roughness index according to e.g.

$$r_i = \sum_{j=2}^{p-1} \left(\hat{b}_j - \frac{\hat{b}_{j-1} + \hat{b}_{j+1}}{2} \right)_i^2 . \tag{25}$$

3. Add a matrix with small random elements to $\mathbf{W}_i$, i.e. form

$$\mathbf{W}_{i,\text{new}} = \mathbf{W}_i + \alpha \Delta \mathbf{W}, \tag{26}$$

   where $\Delta \mathbf{W} \in \mathbb{R}^{p \times A}$ has random entries chosen from e.g. a normal distribution with mean zero and variance one, and where the step factor α is chosen as e.g. $\alpha = 0.001$, or $\alpha = 0.001 e^{-i/10000}$.

4. Perform e.g. an SVD in order to retain orthonormal columns,

$$\mathbf{W}_{i,\text{new}} = \mathbf{U}_i \mathbf{S}_i \mathbf{V}_i^T = \begin{bmatrix} \mathbf{U}_{i,1} & \mathbf{U}_{i,2} \end{bmatrix} \begin{bmatrix} \mathbf{S}_{i,1} \\ \mathbf{0} \end{bmatrix} \mathbf{V}_i^T = \mathbf{U}_{i,1} \mathbf{S}_{i,1} \mathbf{V}_i^T, \tag{27}$$

   and find a randomly modified and orthonormal loading weight matrix

$$\mathbf{W}_{i,\text{mod.}} = \mathbf{U}_{i,1} \in \mathbb{R}^{p \times A}. \tag{28}$$

5. Find again the corresponding predictor $\hat{\mathbf{b}}_{i,\text{mod.}}$ according to (5), the $\text{RMSEC}_{i,\text{mod.}}$ value according to (24) and the roughness index $r_{i,\text{mod.}}$ according to (25). If both the error of calibration RMSEC and the roughness index r decrease, set $\mathbf{W}_{i+1} = \mathbf{W}_{i,\text{mod.}}$, $r_{i+1} = r_{i,\text{mod.}}$ and $\mathbf{b}_{i+1} = \hat{\mathbf{b}}_{i,\text{mod.}}$. Otherwise keep the old values.

6. Compute the RMSEP_{i+1} value using an independent data set, and go to step 8 when no more progress towards a reduced RMSEP is achieved over some iteration steps.

7. Let $i \leftarrow i + 1$ and go to step 3.

8. Set $\hat{\mathbf{W}}_{\text{opt.}} = \mathbf{W}_{i,\text{mod.}}$ and interrupt the search.

Note that the RMSEP value is used for the interrupt decision in step 6 only, i.e. the search as such is based on the modeling set exclusively. Also note that the resulting loading weight matrix $\hat{\mathbf{W}}_{\text{opt.}}$ and the corresponding non-orthogonal score matrix $\hat{\mathbf{T}}_{\text{opt.}} = \mathbf{X}\hat{\mathbf{W}}_{\text{opt.}}$ may be used for interpretational purposes as in the ordinary PLSR case. Alternatively, the factorization (23) with the orthogonal score matrix $\hat{\mathbf{T}}_{\perp}$ may be used for this purpose.

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. 1. Note the overlapping peaks and considerable **X**-noise.

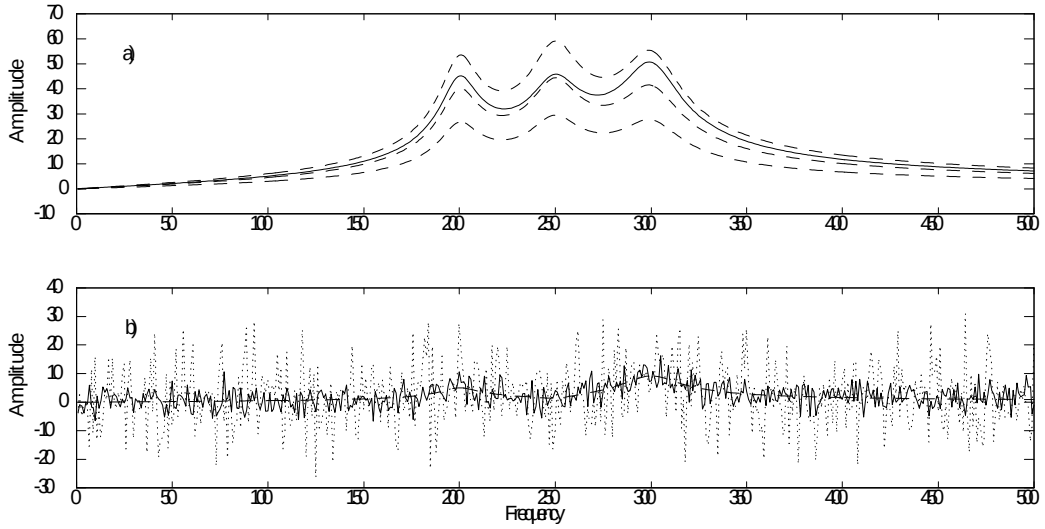

Figure 1. Mean spectrum and standard deviations (Fig. a - dashed lines) plus a typical realization of a noise free original spectrum (Fig. a - solid line), and the corresponding centered and noise corrupted spectrum (Fig. b) of three chemical constituents. The **X**-noise variances are here $r_{ee} = 10$ (solid line) and $r_{ee} = 100$ (dotted line) (see explanation of r_{ee} below). 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 $0 < f \leq 500$ frequency units (f.u.),

$$\begin{aligned} x_k(f) \quad = \quad & \frac{f_1 f}{\sqrt{\left(f_1^2 - f^2\right)^2 + \left(2\zeta f_1 f\right)^2}}(3 + z_{1,k}) + \frac{f_2 f}{\sqrt{\left(f_2^2 - f^2\right)^2 + \left(2\zeta f_2 f\right)^2}}(3 + z_{2,k}) \\ & + \frac{f_3 f}{\sqrt{\left(f_3^2 - f^2\right)^2 + \left(2\zeta f_3 f\right)^2}}(3 + z_{3,k}) + e_k(f) = 3\mathbf{C}_2(f)\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T + \mathbf{C}_2(f)\mathbf{z}_k + e_k(f), \end{aligned} \tag{29}$$

with resonance frequencies $f_1 = 200$ f.u., $f_2 = 250$ f.u., $f_3 = 300$ f.u. and relative dampings $\zeta_1 = \zeta_2 = \zeta_3 = 0.05$, and with $\mathbf{C}_2(f) \in \mathbb{R}^{1\times 3}$. It is also assumed that the variations in the concentration of Constituent 1, Constituent 2 and Constituent 3, denoted $z_{1,k}$, $z_{2,k}$ and $z_{3,k}$, are randomly generated zero mean numbers with normal distributions and variances $r_{zz} = Ez_{1,k}^2 = Ez_{2,k}^2 = Ez_{3,k}^2 = 1$. The noise terms $e_k(f)$ are randomly generated zero mean numbers with normal distribution and equal variances $r_{ee} = Ee_k^2(f)$. Several r_{ee} values were used in the simulations.

It was assumed a scalar response

$$y_k = z_{2,k} = \mathbf{C}_1\mathbf{z}_k + f_k, \tag{30}$$

with $\mathbf{C}_1 = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$ and $Ef_k^2 = 0.0001$. In a practical case this would mean that the primary response of interest would be the concentration of Constituent 2, while the other constituents would be treated as interferants.

Collecting y_k and $\mathbf{x}_k^T$ over $k = 1, 2, \ldots, N$ modeling observations the total model with mean centered data is then

$$\mathbf{y} = \mathbf{Z}\mathbf{C}_1^T + \mathbf{f} \tag{31}$$
$$\mathbf{X} = \mathbf{Z}\mathbf{C}_2^T + \mathbf{E}, \tag{32}$$

where $\mathbf{C}_2 \in \mathbb{R}^{500\times 3}$. Note that this model may be transformed to the model (3,4) by a similarity transformation [10]. With this model the Kalman predictor (16) is replaced by

$$\mathbf{b}_{\mathrm{KF}} = \left(\mathbf{C}_2\mathbf{R}_z\mathbf{C}_2^T + \mathbf{R}_e\right)^{-1}\mathbf{C}_2\mathbf{R}_z\mathbf{C}_1^T, \tag{33}$$

where $\mathbf{R}_z = E\mathbf{z}_k\mathbf{z}_k^T = \mathbf{I}_3$ and $\mathbf{R}_e = E\mathbf{e}_k\mathbf{e}_k^T = r_{ee}\mathbf{I}_{500}$.

PCR and PLSR prediction ability

PCR and PLSR validation results for $M = 100$ Monte Carlo runs at different $\mathbf{X}$-noise levels r_{ee} and with different numbers N of modeling observations using $A = 3$ components and independent validation sets with $N_{\mathrm{val.}} = 1000$ observations are shown in Fig. 2. Here are included also results using the Kalman predictor (33). The RMSEP values at $A = 0$ components were RMSEP = 1.0. Not surprisingly, the predictors deteriorate for small values of N, especially at high noise levels. Note that the difference between PCR and PLSR is more pronounced at high noise levels, and that for large values of N the predictions apparently approach the theoretical Kalman predictions.

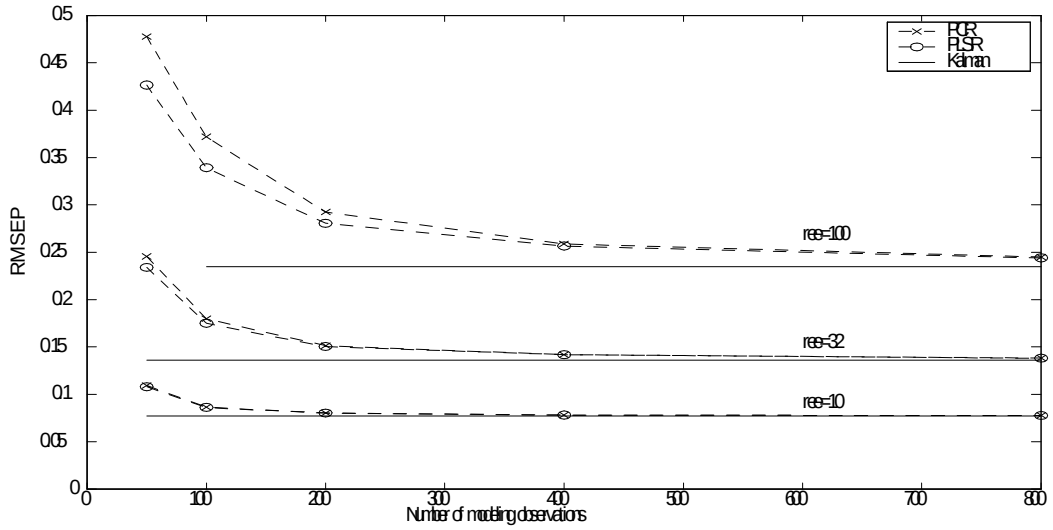

Figure 2. Mean validation PCR and PLSR results for $M = 100$ Monte Carlo runs with $r_{ee} = 10$, 32 and 100, $A = 3$ and different numbers N of modeling observations. The validation results were based on $N_{\mathrm{val.}} = 1000$ observations. The theoretical Kalman predictor results are shown by solid lines.

Fig. 2 indicates that not much is to be gained from optimization at relatively low noise levels combined with relatively long modeling data (e.g. $r_{ee} = 10$ and $N = 200$).

Optimization results

An optimization result for $A = 3$ components using the algorithm in Section 3 with a step factor $\alpha = 0.001e^{-i/10000}$ is shown in Fig. 3. The $\mathbf{X}$-noise level was here $r_{ee} = 10$ (see Fig. 1 and 2), while the

number of observations in the modeling and validation data were $N = 50$ and $N_{\text{val.}} = 1000$ respectively. The theoretical RMSEK value based on the Kalman predictor (33), as well as the corresponding LS predictor (18) are also shown. The relative RMSEP reduction is 20 %, while the theoretically possible reduction is at most 25 %. The absolute reduction is 2.0 %, which could be compared with the standard deviation of 1 % for the y_k measurements. However, note that the variance in the y_k observations gives a 1 % contribution to all RMSE values, and therefore is of no importance for the differences and thus the optimization improvement. $A = 3$ components gave the best predictor both before and after the optimization. Also note that the RMSEP value is plotted for illustration of the search progress only, while the optimization as such was based on RMSEC exclusively.

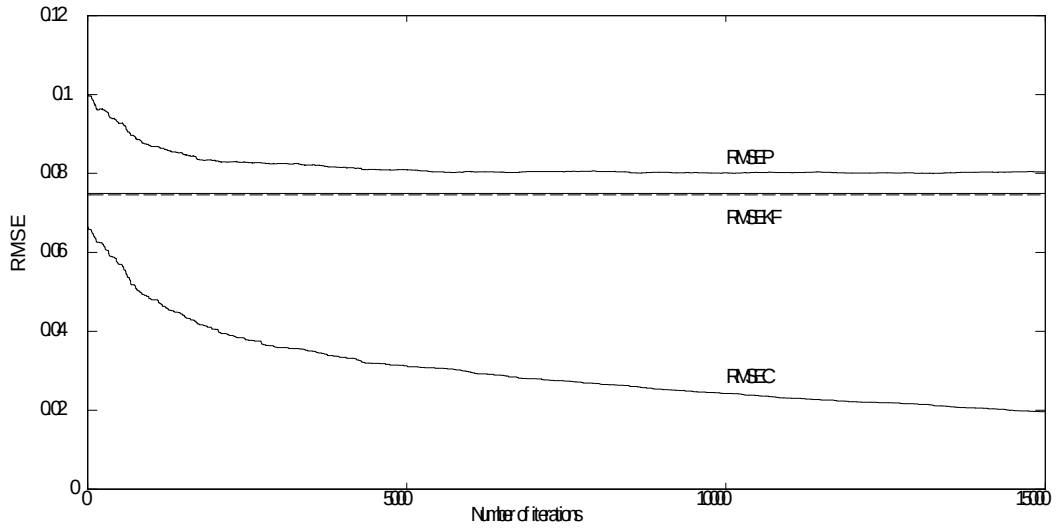

Figure 3. Optimization results for simulation with $r_{ee} = 10$, $N = 50$, $A = 3$ and $N_{\text{val.}} = 1000$, using a step factor $\alpha = 0.001e^{-i/10000}$. The relative RMSEP reduction is 20 % of a theoretically possible 25 %. The theoretical validation result using the Kalman predictor (33) is shown by dashed straight line, while the corresponding LS result according to (18) is shown as solid straight line.

The search for an optimal predictor was interrupted after 15000 iterations (ca. 18 min. on a Pentium 4 PC). The roughness index (25) was reduced from an initial value $r_{\text{init.}} = 9.8\ 10^{-5}$ to a final value $r_{\text{final}} = 4.6\ 10^{-5}$. The noisy variations in $\hat{\mathbf{b}}$ had approximately the same amplitude in the initial and the final predictor, while the variations in the final predictor had a clearly reduced content of "rapid" fluctuations.

A typical optimization result with the $\mathbf{X}$-noise level and the number of modeling observations increased to $r_{ee} = 100$ and $N = 100$ is shown in Fig. 4 (see also Fig. 1). Here a fixed step factor $\alpha = 0.0001$ was used. Tests with a random α (unity distribution between 0 and 0.001), and with an exponentially declining $\alpha = 0.001e^{-i/10000}$ gave somewhat faster convergences, but no improvement of the final result. Some simulations gave in fact a slowly increasing RMSEP value at very high numbers of iterations.

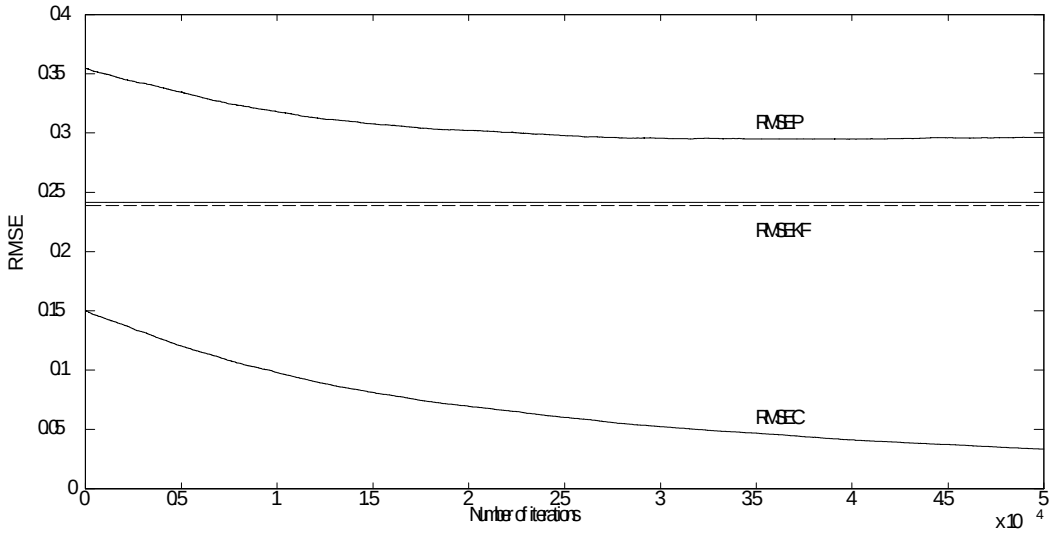

Figure 4. Optimization results for simulation with $r_{ee} = 100$, $N = 100$, $A = 3$ and $N_{\text{val.}} = 1000$, using a step factor $\alpha = 0.0001$. The relative RMSEP reduction is here 17 % of a possible 32 %. The theoretical validation result using the Kalman predictor (33) is shown by dashed straight line, while the corresponding LS result according to (18) is shown as solid straight line.

5 Metal ion mixtures example

The optimization method developed in Section 3 was tested on a data set made available from the Wentzell Group [11]. The data set, labeled "inortrun", was "obtained through a carefully designed experiment involving three-component mixtures of metal ions (Co(II), Cr(III), Ni(II))". The $\mathbf{X}$ measurements were absorbances at $p = 151$ frequencies, while the concentration of Co was used as the response variable y, and the total number of observations is $N_{\text{total}} = 128$. Based on the mixture preparation methods, the uncertainty in the y_k values can be assumed to be less than 1 % [12]. The $N = 25$ observations number 31 to 55 were used for modeling, while the other $N_{\text{val.}} = 103$ observations were used for validation. The data were autoscaled, and the optimization result using $A = 3$ components is shown in Fig. 5. The relative RMSEP reduction is 53 %. Again note that the RMSEP value is not used in the optimization algorithm as such. The search for an optimal predictor was here interrupted after 50000 iterations (ca. 3 min. on a Pentium 4 PC). Note that in this real world data case a theoretical Kalman predictor result as in Fig. 3 and 4 is not available.

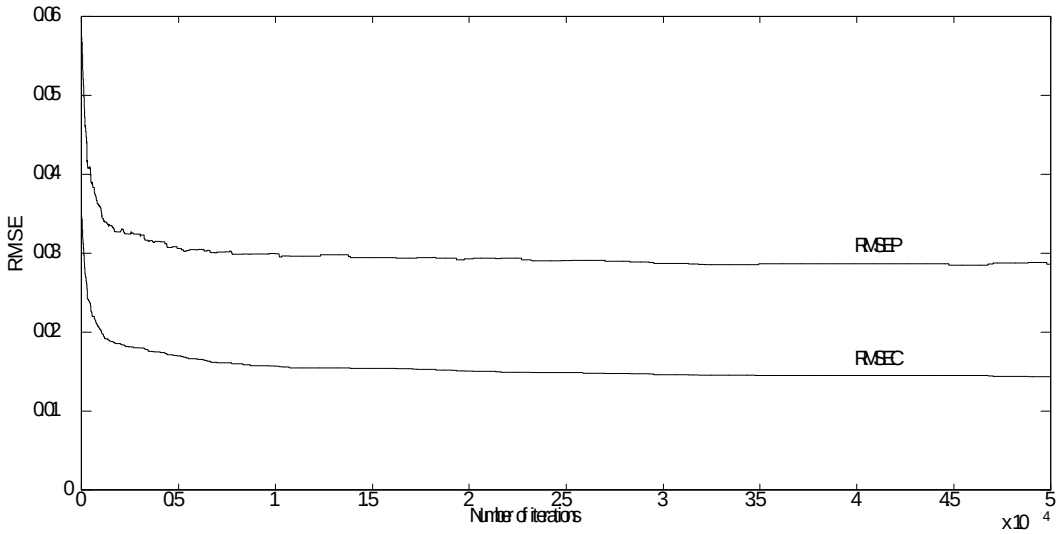

Figure 5. Optimization results for metal ion mixture with $A = 3$ components and step factor $\alpha = 0.001$. The relative RMSEP reduction is 53 %.

The RMSEP values for different numbers of components are shown in Fig. 6, now based on 10000 iterations. Due to the low number of modeling observations ordinary PLSR gave a minimum for $A = 4$ components, and compared to that the optimization gave a 22 % RMSEP reduction, and at the same time a reduction to $A = 3$ components. This corresponds to an absolute improvement of 0.6 %, while the uncertainty in the y_k values can be assumed to be less than 1 %. Also here the y_k uncertainty affects both RMSEP and $\text{RMSEP}_{\text{opt.}}$, such that the difference is not affected. However, the most important improvement might be the reduction in number of components, and after the optimization also a model with $A = 2$ components might be considered.

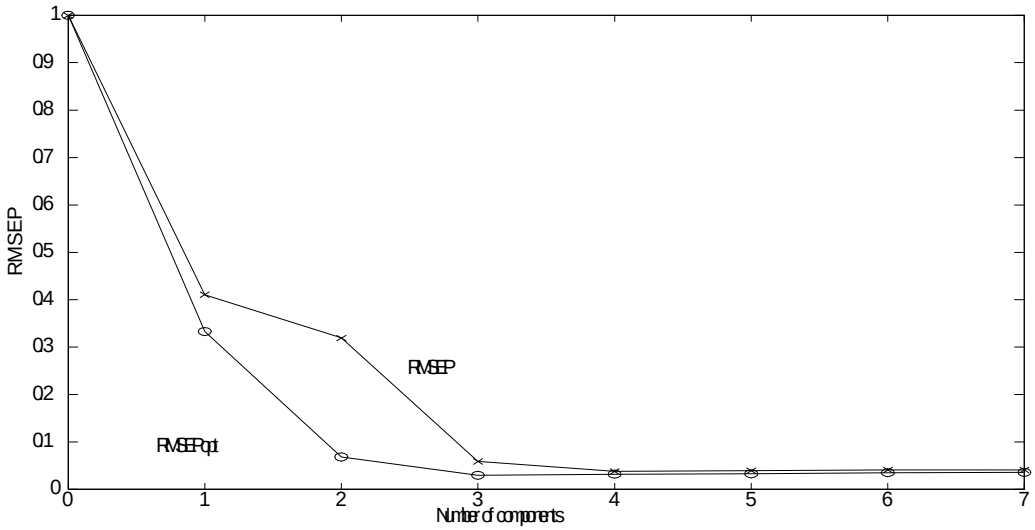

Figure 6. Prediction RMSEP values for metal ion mixture at different numbers of components before and after optimization using 10000 iterations.

6 Summary and conclusions

The theoretical basis for the numerical optimization algorithm presented in Section 3 is the fact that under the assumption of a linear LV model an optimal loading weight matrix can be shown to exist, resulting in the best linear unbiased estimator (BLUE). Also assuming normal LV and **X**-noise distributions, the BLUE is the best estimator whatsoever. This basis is summarized in Theorem 1. However, in order to prevent from overfitting to the modeling data, the numerical search for the optimal loading weight matrix must be constrained in some way or another. In the present algorithm this is done in an *ad hoc* fashion, by requiring that a certain roughness index for the resulting predictor is not increasing. It is possible that other types of constraints may be used, and they are indeed needed for cases where a smooth predictor cannot be expected.

The optimization algorithm is at present very simple, and could possibly be improved by further work. It must be remembered though, that assuming a useful step factor the optimization need to be done only once, and the computational demand may thus not be of critical importance. However, practical step factor guidelines remain to be worked out. The algorithm is also restricted to handling of a scalar response variable, although this may be applied separately for each of several responses.

The simulation example indicates that considerable relative prediction improvements may be obtained (20 % and 17 % in two cases with different noise levels), and this is confirmed in the metal ion mixtures example (22 % combined with a reduction of number of components). The simulation example also shows, however, that little is to be gained at a combination of relatively low noise levels and relatively long modeling data.

References

[1] Burnham, A.J, Viveros R, MacGregor J.F., Frameworks for Latent Variable Multivariate Regression, *J. Chemometrics* 1996;**10**:31-45.

[2] Helland, I.S., On the structure of partial least squares regression, *Communications in statistics* 1988;**17**:581-607.

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

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

[5] Isaksson, T, Næs T., Selection of Samples for Calibration in Near-Infrared Spectroscopy. Part II: Selection Based on Spectral Measurements, *Applied Spectroscopy* 1990;**44**:1152-1158.

[6] Grewal, M.S, Andrews A.P., *Kalman Filtering: Theory and Practice*, Prentice Hall: New Jersey, 1993.

[7] Gelb, A., *Applied Optimal Estimation*, MIT Press, Mass., 1974.

[8] Berntsen, H., Utvidet Kalmanfilter og multivariabel kalibrering, *Report STF48 A88019*, SINTEF, Trondheim, Norway, 1988.

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

[10] Kailath, T., *Linear Systems*, Prentice Hall: New Jersey, 1980.

[11] http://www.dal.ca/~pdwentze/download.htm

[12] Wentzell, P., Personal communication, 2002 .