Dynamic system multivariate calibration with low-sampling-rate y data

Rolf Ergon* and Maths Halstensen

Telemark Institute of Technology, Porsgrunn, Norway

SUMMARY

When the data in principal component regression (PCR) or partial least squares regression (PLSR) form time series, it may be possible to improve the prediction/estimation results by utilizing the correlation between neighboring observations. The estimators may then be identified from experimental data using system identification methods. This is possible also in cases where the response variables in the experimental data are sampled at a low and possibly irregular rate, while the regressor variables are sampled at a higher rate. After a discussion of the options available, the paper shows how the autocorrelation of the regressor variables in such multirate sampling cases may be utilized by identification of parsimonious output error (OE) estimators. An example using acoustic power spectrum regressor data is finally presented.

KEY WORDS: dynamic; multirate; multivariate; calibration

1. PROBLEM

When the **X** and **y** data (assuming a single response variable) in ordinary PCR (principal component regression) or PLSR (partial least squares regression) form time series, it may be possible to improve the prediction/estimation results by utilizing the correlation between neighboring observations. After compression of the **X** data into a latent variables matrix $\hat{\mathbf{T}}$, this can basically be done by identification of an FIR (finite impulse response), ARMAX (autoregressive moving average with extra inputs), ARX (autoregressive with extra inputs) or OE (output error) type of model. The estimator may thus be found from experimental data by a combination of multivariate calibration [1] and system identification [2] methods.

In many cases of practical interest it is possible to obtain experimental **X** data at a high sampling rate, while the experimental **y** data for practical and economical reasons are obtained at a lower and possibly also irregular rate. The problem of the present paper is to find an identification method that handles such multivariate and multirate sampling time series data.

After an introductory discussion of the options available, a solution to the problem using an OE model structure is presented. An example using acoustic power spectrum **X** data is finally presented.

DISCUSSION

2.1. FIR models

Assuming a dynamic and asymptotically stable multi-input–single-output system with a known input sequence $\mathbf{u}_k$ and a recorded response sequence y_k, the FIR model is

$$y_k \approx \sum_{i=0}^{L} \mathbf{h}_i^{\mathrm{T}} \mathbf{u}_{k-i} + \eta_k \tag{1}$$

where L is chosen sufficiently large in order to capture the essential dynamics of the system, and where η_k is a non-white stochastic sequence. This means that LT must be larger than the settling time of the system, where T is the sampling interval for the input variables. We can then determine estimates $\hat{\mathbf{h}}_0, \hat{\mathbf{h}}_1, \ldots, \hat{\mathbf{h}}_L$ of the Markov parameters as a least squares (LS) solution of (1). Such a system identification is possible also when most of the y_k values are missing, in which case we use only the y_j samples that are available, together with the corresponding past and present $\mathbf{u}_k$ values.

There are two principal drawbacks related to the identification of FIR models.

- The models found will be biased owing to both the truncation and the lack of noise modeling. This might, however, be a minor problem in a practical case without a well-defined underlying system. The problem is also reduced when the output is sampled at a low rate, since the noise terms η_j will then not be consecutive.
- Since LT must be longer than the settling time of the system, a large number of Markov parameters may have to be identified, and this requires a correspondingly large number of response observations. This is a serious drawback, especially since it often is a part of the problem that the number of response observations is very limited. Also note that the dimension of each Markov parameter $\mathbf{h}_i$ is determined by the number of input and output variables.

Some authors have discussed the use of PCR and PLSR in order to determine FIR models for dynamic systems. Ricker [3] discussed identification of FIR models using PLSR and singular value decomposition (SVD), and applied this to an anaerobic wastewater treatment plant. MacGregor *et al* [4] looked at several methods for the identification of dynamic models by use of PLSR. Wise and Ricker [5] discussed identification of FIR models by use of PCR, and looked especially at frequency response properties. Dayal and MacGregor [6] focused on robustness issues.

2.2. ARMAX models

Assuming a multiple-input–single-output system, an ARMAX model would be of the type

$$\begin{aligned} y_k = & -a_1 y_{k-1} - a_2 y_{k-2} - \ldots + \mathbf{b}_0^{\mathrm{T}} \mathbf{u}_k + \mathbf{b}_1^{\mathrm{T}} \mathbf{u}_{k-1} + \mathbf{b}_2^{\mathrm{T}} \mathbf{u}_{k-2} + \ldots \\ & + c_0 e_k + c_1 e_{k-1} + c_2 e_{k-2} + \ldots \end{aligned} \tag{2}$$

where e_k is a white noise sequence. More generally we would have

$$\mathbf{A}(q^{-1}) y_k = \mathbf{B}(q^{-1}) \mathbf{u}_k + \mathbf{C}(q^{-1}) e_k \tag{3}$$

where $\mathbf{A}(q^{-1})$, $\mathbf{B}(q^{-1})$ and $\mathbf{C}(q^{-1})$ are polynomials in the unit time delay operator q^{-1}.

There are two principal drawbacks related to the identification of ARMAX estimators.

- They make use of past y_k data and thus cannot be identified when y is sampled at a lower rate than $\mathbf{u}$.

- They are in any case non-optimal when also secondary system outputs are used as inputs to the estimator, i.e. when we are identifying a model

$$\mathbf{A}(q^{-1})y_{1,k} = \mathbf{B}_1(q^{-1})\mathbf{u}_k + \mathbf{B}_2(q^{-1})\mathbf{y}_{2,k} + \mathbf{C}(q^{-1})e_k \tag{4}$$

  The reason for this is that past $y_{1,k}$ values will not be available in the estimation stage, and an ARMAX estimator will then not utilize the $\mathbf{y}_2$ information in an optimal way [7].

2.3. *ARX models*

ARX models are ARMAX models without moving average (MA) noise modeling, i.e. of the structure given in (3) with $\mathbf{C}(q^{-1}) = c_0$, and the parameters may thus be determined by an ordinary LS solution. ARX estimators

- make use of past y_k data and thus cannot be identified when y is sampled at a lower rate than $\mathbf{u}$;
- are in any case non-optimal owing to the lack of MA noise modeling;
- are even more non-optimal when also secondary system outputs are used as inputs to the estimator, for the same reason as for the ARMAX estimators discussed above.

Several authors have discussed the use of PCR and PLSR in order to determine ARX models for dynamic systems. Wise [8] studied the relations between ARX models and PCA based on lagged values of both the system inputs and the system outputs. Qin and McAvoy [9] used PLSR in a similar way to find an ARX model for a catalytic reforming system. Wise *et al* [10] compared neural networks, PLSR and a genetic algorithm used for identification of non-linear FIR and ARX models. Dayal and MacGregor [11] presented a recursive PLSR algorithm and identified an ARX model used for adaptive control of a simulated stirred tank reactor. Wikström *et al* [12] used PLSR for time series modeling related to an electrolysis process (they had no known inputs and were thus identifying an AR model). Harnett *et al* [13] extended the work of Wise [8] in order to facilitate the development of a predictive model of the overheads condenser and reflux drum system for a distillation column.

2.4. *OE models*

OE estimators

- can be identified also when most of the ordinary y observations are missing (see theoretical development below);
- use a small number of parameters;
- are optimal in the sense that they are unbiased and have minimized variances also when secondary system outputs are used as estimator inputs [7].

3. THEORETICAL OE ESTIMATOR

Assume the known discrete-time state space plant model

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k + \mathbf{G}\mathbf{v}_k \\ y_{1,k} &= \mathbf{C}_1\mathbf{x}_k + \mathbf{D}_1\mathbf{u}_k + w_{1,k} \\ \mathbf{y}_{2,k} &= \mathbf{C}_2\mathbf{x}_k + \mathbf{D}_2\mathbf{u}_k + \mathbf{w}_{2,k} \end{aligned} \tag{5}$$

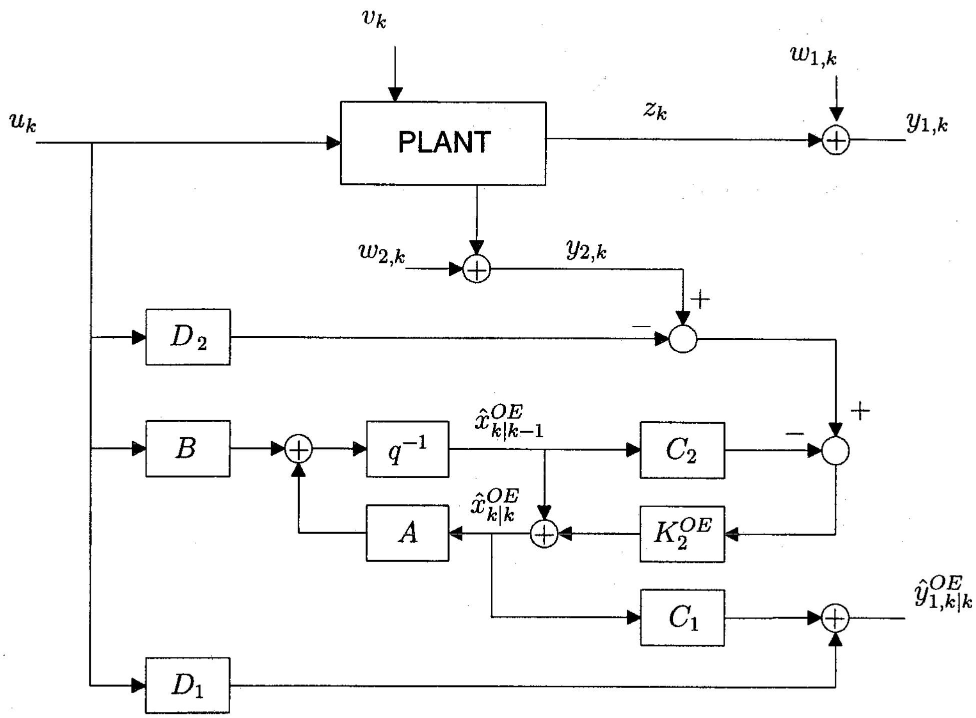

Figure 1. Optimal primary output estimator.

where $\mathbf{x}_k$ is the state vector. Here $y_{1,k}$ are the primary response observations that in ordinary PCR/PLSR with a single response variable would be collected in the $\mathbf{y}$ vector, while the known inputs $\mathbf{u}_k$ and the secondary outputs $\mathbf{y}_{2,k}$ would be collected in the $\mathbf{X}$ matrix. The process and measurement noise sequences $\mathbf{v}_k$, $w_{1,k}$ and $\mathbf{w}_{2,k}$ are assumed to be independent and white.

Assuming the system matrices known, the optimal y_1 estimator is then obtained by the Kalman filter in Figure 1.

The matrix $\mathbf{K}_2^{\mathrm{OE}}$ is here a Kalman gain found from the algebraic Riccati equation

$$\mathbf{P}^{\mathrm{OE}} = \mathbf{A}\mathbf{P}^{\mathrm{OE}}\mathbf{A}^{\mathrm{T}} + \mathbf{G}\mathbf{R}_{\mathrm{v}}\mathbf{G}^{\mathrm{T}} - \mathbf{A}\mathbf{P}^{\mathrm{OE}}\mathbf{C}_2^{\mathrm{T}}(\mathbf{C}_2\mathbf{P}^{\mathrm{OE}}\mathbf{C}_2^{\mathrm{T}} + \mathbf{R}_{22})^{-1}\mathbf{C}_2\mathbf{P}^{\mathrm{OE}}\mathbf{A}^{\mathrm{T}} \tag{6}$$

and

$$\mathbf{K}_2^{\mathrm{OE}} = \mathbf{P}^{\mathrm{OE}}\mathbf{C}_2^{\mathrm{T}}(\mathbf{C}_2\mathbf{P}^{\mathrm{OE}}\mathbf{C}_2^{\mathrm{T}} + \mathbf{R}_{22})^{-1} \tag{7}$$

where $\mathbf{P}^{\mathrm{OE}} = E(\mathbf{x}_k - \hat{\mathbf{x}}^{\mathrm{OE}}_{k|k-1})(\mathbf{x}_k - \hat{\mathbf{x}}^{\mathrm{OE}}_{k|k-1})^{\mathrm{T}}$ is the minimized prediction state estimation covariance [14].

The optimal estimator given by the block diagram in Figure 1 is

$$\begin{aligned}\hat{y}^{\mathrm{OE}}_{1,k|k} &= \mathbf{C}_1(\mathbf{I} - \mathbf{K}_2^{\mathrm{OE}}\mathbf{C}_2)(q\mathbf{I} - \mathbf{A} + \mathbf{A}\mathbf{K}_2^{\mathrm{OE}}\mathbf{C}_2)^{-1} \\ &\quad \times [(\mathbf{B} - \mathbf{A}\mathbf{K}_2^{\mathrm{OE}}\mathbf{D}_2)\mathbf{u}_k + \mathbf{A}\mathbf{K}_2^{\mathrm{OE}}\mathbf{y}_{2,k}] + \mathbf{C}_1\mathbf{K}_2^{\mathrm{OE}}(\mathbf{y}_{2,k} - \mathbf{D}_2\mathbf{u}_k) + \mathbf{D}_1\mathbf{u}_k\end{aligned} \tag{8}$$

This estimator corresponds to the OE model

$$y_{1,k} = \frac{\mathbf{B}_1(q^{-1})\mathbf{u}_k + \mathbf{B}_2(q^{-1})\mathbf{y}_{2,k}}{\mathbf{A}(q^{-1})} + \vartheta_k \tag{9}$$

where $\mathbf{A}\,(q^{-1})$, $\mathbf{B}_1\,(q^{-1})$ and $\mathbf{B}_2\,(q^{-1})$ are polynomials in the unit time delay operator q^{-1}, and ϑ_k is a non-white stochastic process.

Note that for a first-order system with $\mathbf{C}_1 = 1$ and $\mathbf{u}_k = 0$, Equation (8) can be simplified to

$$\hat{y}^{\mathrm{OE}}_{1,k|k} = [1 - (A - \mathrm{A}\mathbf{K}_2^{\mathrm{OE}}\mathbf{C}_2)q^{-1}]^{-1}\mathbf{K}_2^{\mathrm{OE}}\mathbf{y}_{2,k} \tag{10}$$

4. ESTIMATOR IDENTIFICATION

The block diagram in Figure 1 may be rearranged as shown in Figure 2, where the notation $(\cdot)^{\mathrm{pred}}$ indicates parameter values that are not necessarily correct (although some useful initial values are needed). Figure 2 illustrates the prediction error method (PEM) for system identification, adapted to minimize the error in the current estimate $y^{\mathrm{pred}}_{1,k|k}$.

As indicated in the figure, the parameter values are successively modified until a criterion function

$$V_N(\theta) = \frac{1}{N}\sum_{k=1}^{N}\varepsilon^2_{1,k}(\theta) \tag{11}$$

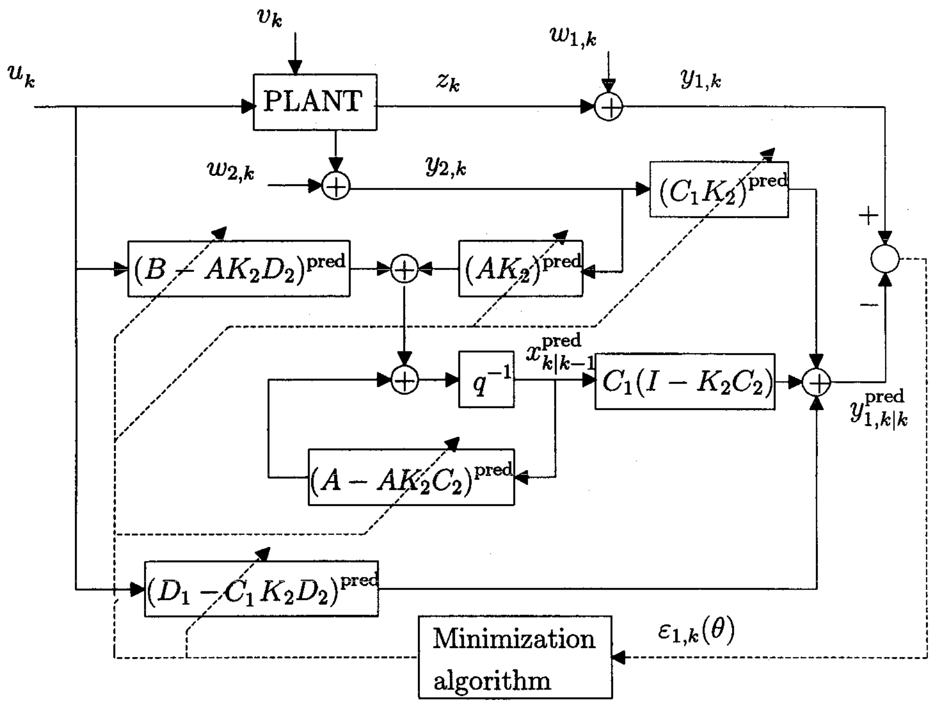

Figure 2. Identification by estimation error minimization.

is minimized. Here

$$\varepsilon_{1,k}(\theta) = y_{1,k} - y_{1,k|k}^{\mathrm{pred}} \tag{12}$$

is the estimation error and N is the number of y_1 observations [15–18].

When not all $y_{1,k}$ values are available, only the estimation errors $\varepsilon_{1,j}$ that can be computed may be used [19]. The most difficult part may in fact be to find useful initial parameter values for the minimization, and several *ad hoc* approaches can be used to solve that problem [18,19].

Minimization of (11) can be performed only after a specification of the structure and model order, and the normal procedure is then to select a chain of canonical structures with increasing model order, starting with a low-order model [2]. An alternative would be to find the model order by use of a subspace identification method [20], which would give an ARMAX type of model. However, this is not a feasible choice in the case of a low primary output sampling rate.

5. NUMERICAL ALGORITHM

The practical identification in the multirate sampling case can be performed by a modified version of the *pem. m* function in the System Identification Toolbox for use with Matlab, which uses an iterative Gauss–Newton algorithm [21]. The modifications are necessary in order to handle the situation when only some of the $y_{1,k}$ observations are available.

6. PRINCIPAL COMPONENT SOLUTION

With a large number of secondary $\mathbf{y}_2$ variables, the number of parameters to identify will be correspondingly large. In many cases, fortunately, the $\mathbf{y}_2$ variables are highly collinear, and the $\mathbf{Y}_2$ data may then be compressed by use of principal component analysis (PCA) and the factorization [1,22]

$$\mathbf{Y}_2 = \hat{\mathbf{T}}\hat{\mathbf{P}}^{\mathrm{T}} + \mathbf{E} \tag{13}$$

Using estimated latent variables

$$\hat{\boldsymbol{\tau}}_k = \hat{\mathbf{P}}^{\mathrm{T}}\mathbf{y}_{2,k} \tag{14}$$

instead of $\mathbf{y}_{2,k}$, the estimator (8) is replaced by

$$\begin{aligned} \hat{y}_{1,k|k}^{\mathrm{PCA+OE}} &= \mathbf{C}_1(\mathbf{I} - \mathbf{K}_{\mathrm{P}}\hat{\mathbf{P}}^{\mathrm{T}}\mathbf{C}_2)(q\mathbf{I} - \mathbf{A} + \mathbf{A}\mathbf{K}_{\mathrm{P}}\hat{\mathbf{P}}^{\mathrm{T}}\mathbf{C}_2)^{-1} \\ &\quad \times [(\mathbf{B} - \mathbf{A}\mathbf{K}_{\mathrm{P}}\hat{\mathbf{P}}^{\mathrm{T}}\mathbf{D}_2)\mathbf{u}_k + \mathbf{A}\mathbf{K}_{\mathrm{P}}\hat{\boldsymbol{\tau}}_k] + \mathbf{C}_1\mathbf{K}_{\mathrm{P}}(\hat{\boldsymbol{\tau}}_k - \hat{\mathbf{P}}^{\mathrm{T}}\mathbf{D}_2\mathbf{u}_k) + \mathbf{D}_1\mathbf{u}_k \end{aligned} \tag{15}$$

where the Kalman gain is given by appropriately modified versions of (6) and (7), i.e.

$$\mathbf{K}_{\mathrm{P}} = \mathbf{P}_{\mathrm{P}}\mathbf{C}_2^{\mathrm{T}}\hat{\mathbf{P}}(\hat{\mathbf{P}}^{\mathrm{T}}\mathbf{C}_2\mathbf{P}_{\mathrm{P}}\mathbf{C}_2^{\mathrm{T}}\hat{\mathbf{P}} + \hat{\mathbf{P}}^{\mathrm{T}}\mathbf{R}_{22}\hat{\mathbf{P}})^{-1} \tag{16}$$

with $\mathbf{P}_{\mathrm{P}}$ determined by the algebraic Riccati equation

$$\mathbf{P}_{\mathrm{P}} = \mathbf{A}\mathbf{P}_{\mathrm{P}}\mathbf{A}^{\mathrm{T}} + \mathbf{G}\mathbf{R}_{\mathrm{v}}\mathbf{G}^{\mathrm{T}} - \mathbf{A}\mathbf{P}_{\mathrm{P}}\mathbf{C}_2^{\mathrm{T}}\hat{\mathbf{P}}(\hat{\mathbf{P}}^{\mathrm{T}}\mathbf{C}_2\mathbf{P}_{\mathrm{P}}\mathbf{C}_2^{\mathrm{T}}\hat{\mathbf{P}} + \hat{\mathbf{P}}^{\mathrm{T}}\mathbf{R}_{22}\hat{\mathbf{P}})^{-1}\hat{\mathbf{P}}^{\mathrm{T}}\mathbf{C}_2\mathbf{P}_{\mathrm{P}}\mathbf{A}^{\mathrm{T}} \tag{17}$$

The identification of this estimator is performed as indicated in Figure 2, only now the $\mathbf{y}_{2,k}$ data are replaced by the $\hat{\boldsymbol{\tau}}_k$ data as given by (14). For the special case of a first-order estimator, Equation (10) is then modified to

$$\hat{y}_{1,k|k}^{\mathrm{PCA+OE}} = [1 - (A - A\mathbf{K}_{\mathrm{p}}\hat{\mathbf{P}}^{\mathrm{T}}\mathbf{C}_2)q^{-1}]^{-1}\mathbf{K}_{\mathrm{p}}\hat{\boldsymbol{\tau}}_k \tag{18}$$

7. ACOUSTIC DATA EXAMPLE

7.1. *Introduction*

Acoustic chemometrics is based on signals from an acoustic sensor (accelerometer) placed e.g. on, or slightly downstream of, a standard orifice plate. Observations of the power spectrum of the sensor signal are collected in the $\mathbf{X} = \mathbf{Y}_2$ matrix and calibrated against physical $y = y_1$ primary quantities such as multicomponent mixture concentrations, density, etc. using e.g. a standard PLSR method [23]. The information in the large number of collinear $\mathbf{Y}_2$ variables is then compressed into a small number of latent variables $\hat{\boldsymbol{\tau}}_k^{\mathrm{PLSR}} = \hat{\mathbf{W}}^{\mathrm{T}}\mathbf{y}_{2,k}$, where $\hat{\mathbf{W}}$ is the PLSR weighting matrix.

7.2. *Experiment*

In an experiment on a test rig at Telemark Institute of Technology, Porsgrunn, Norway the flow rate of ordinary drinking water was measured by use of an orifice plate. More precisely, the differential pressure across the orifice was measured and used as the response variable y_1, while the acoustic power spectral densities at 1024 frequencies were used as $\mathbf{y}_2$ variables. The sampling interval was 5·3 s and the total number of observations was 200. No known inputs $\mathbf{u}$ were recorded. Six outliers in the acoustic power spectrum results were corrected by use of interpolated values (in a time series analysis, outliers cannot simply be removed).

The response variable had a considerable content of high-frequency components, as shown in Figure 3.

As indicated in the figure, the data set was separated into two parts, one for modeling by system identification and the other for validation.

7.3. *Multirate identification*

In order to test the method presented above, static PLSR and dynamic PCA+OE estimators were identified using only every fifth response observation in the modeling data (the first 20 open circles in Figure 3). We thus consider a multirate system with $\mathbf{y}_2$ sampled with intervals $T_2 = 5{\cdot}3$ s, while y_1 is sampled with intervals $T_1 = 26{\cdot}5$ s. We also consider a difficult case with only $N_1 = 20$ samples of y_1. The estimators were validated against the complete validation set using all 100 primary output samples in the second part of the data set.

Remark. The use of every fifth observation from high-sampling-rate data is of course not optimal. However, if the response variable had been e.g. a concentration of a certain component in a multicomponent mixture, and the measurements had to be done through laboratory analyses, the response sampling rate would very likely for practical and economic reasons have been lower than the obtainable acoustic data sampling rate. Also the validation would in such a case have to be done against low-sampling-rate data.

Both an ordinary PLSR estimator [16,17]

$$\hat{\mathbf{b}}^{\mathrm{PLSR}} = \hat{\mathbf{W}}(\hat{\mathbf{W}}^{\mathrm{T}}\mathbf{Y}_{2,\mathrm{low}}^{\mathrm{T}}\mathbf{Y}_{2,\mathrm{low}}\hat{\mathbf{W}})^{-1}\hat{\mathbf{W}}^{\mathrm{T}}\mathbf{Y}_{2,\mathrm{low}}^{\mathrm{T}}\mathbf{y}_1 = \hat{\mathbf{W}}\hat{\mathbf{b}}_{\mathrm{T}}^{\mathrm{PLSR}} \tag{19}$$

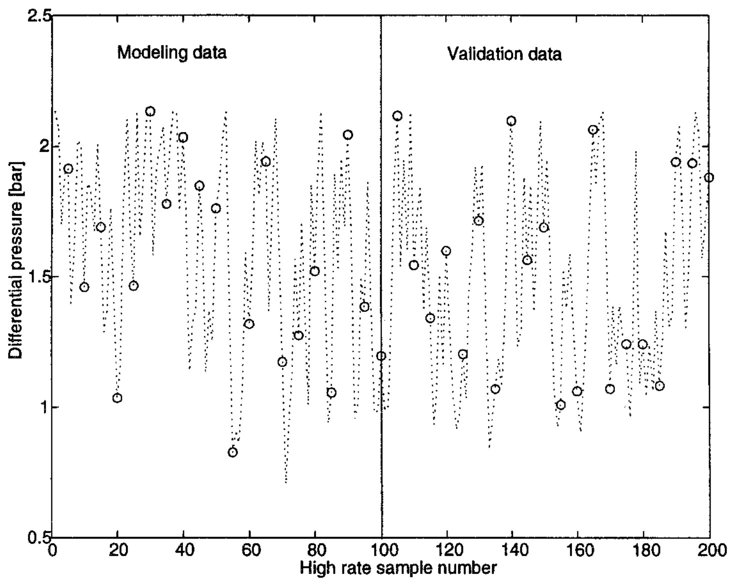

Figure 3. Differential pressure across orifice plate (dotted curve), with low-sampling-rate measurements (open circles).

and the first-order PCA+OE estimator (18) were determined using only the 20 response observations considered available. Here $\mathbf{Y}_{2,\mathrm{low}}$ contains only low-sampling-rate $\mathbf{y}_{2,k}$ data.

7.4. *Number of parameters*

The number of parameters in the PLSR estimator is equal to the number of components a chosen, while the PCA+OE estimator has $n+(n+1)a$ parameters, where n is the model order (i.e. 13 parameters for $a=6$ and $n=1$, and 20 parameters for $a=6$ and $n=2$).

7.5. *Initial values*

As must be expected with the considerable high-frequency content in the response signal, it turned out to be difficult to find useful initial values for the PCA+OE estimators, and use of random values did not work. Instead, the initial parameters in a first-order state space realization

$$\begin{aligned} x_{k+1} &= f x_k + \mathbf{h}^{\mathrm{T}} \hat{\boldsymbol{\tau}}_k \\ y_{1,k} &= x_k + \mathbf{m}^{\mathrm{T}} \hat{\boldsymbol{\tau}}_k \end{aligned} \tag{20}$$

corresponding to the OE estimator

$$\hat{y}_{1,k|k} = \frac{\mathbf{m}^{\mathrm{T}} + (\mathbf{h}^{\mathrm{T}} - f\mathbf{m}^{\mathrm{T}})q^{-1}}{1 - fq^{-1}} \hat{\boldsymbol{\tau}}_k \tag{21}$$

were chosen such that the structure of estimator (18) was obtained. This requires that $f=A-A\mathbf{K}_{\mathrm{P}}\hat{\mathbf{P}}^{\mathrm{T}}\mathbf{C}_2$, $\mathbf{m}^{\mathrm{T}}=\mathbf{K}_{\mathrm{P}}$ and $\mathbf{h}^{\mathrm{T}}=f\mathbf{m}^{\mathrm{T}}$, resulting in the static estimator (with $q^{-1}=1$)

$$\hat{y}_{1,k|k} = \frac{\mathbf{m}^{\mathrm{T}}}{1-f}\hat{\boldsymbol{\tau}}_k = \frac{\mathbf{m}^{\mathrm{T}}}{1-f}\hat{\mathbf{P}}^{\mathrm{T}}\mathbf{y}_{2,k} = (\hat{\mathbf{b}}_{\mathrm{T}}^{\mathrm{PCA}})^{\mathrm{T}}\hat{\mathbf{P}}^{\mathrm{T}}\mathbf{y}_{2,k} \tag{22}$$

A comparison with the PLSR estimator (19) now indicates the initial guesses

$$\mathbf{m}_0^{\mathrm{T}} = (1-f_0)\ (\pm\hat{\mathbf{b}}_{\mathrm{T}}^{\mathrm{PLSR}})^{\mathrm{T}} \tag{23}$$

and

$$\mathbf{h}_0^{\mathrm{T}} = f_0\mathbf{m}_0^{\mathrm{T}} = f_0(1-f_0)\ (\pm\hat{\mathbf{b}}_{\mathrm{T}}^{\mathrm{PLSR}})^{\mathrm{T}} \tag{24}$$

Here the sign indeterminacy is due to the use of sign-indeterminate PCA components. It now only remains to select f_0 and the sign. After a systematic search the initial value was chosen as $f_0 = 0{\cdot}25$, combined with negative sign in (23) and (24). Values in the range $0{\cdot}15 < f_0 < 0{\cdot}30$ gave very similar results, and values in an even wider range could be used when combined with repeated identifications with randomly modified initial values.

7.6. Results

The PCA+OE estimator (21) for different numbers of components found in this way gave the validation RMSEP results shown in Figure 4, where also the validation results for the corresponding ordinary PLSR estimator (19) are shown.

As can be seen in Figure 4, the first-order dynamic estimator gave an approximately 30% reduction of the validation RMSEP value at the optimal number of components $a = 6$.

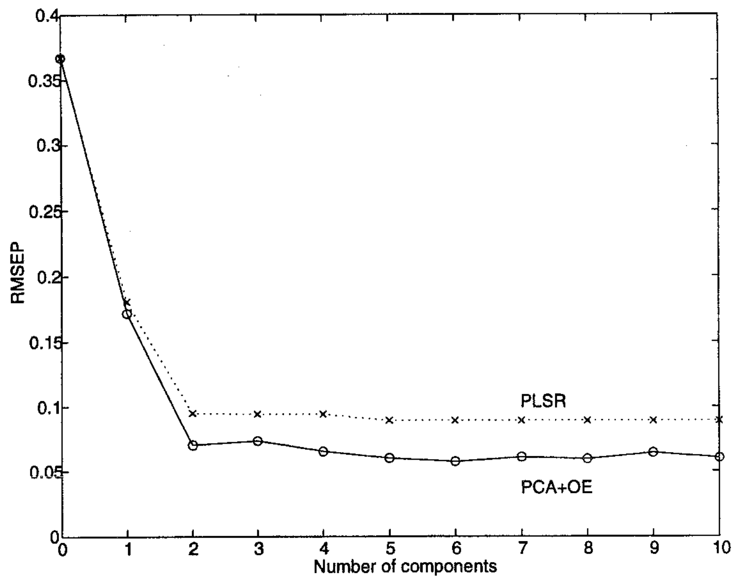

Figure 4. Validation RMSEP results for static PLSR and dynamic PCA+OE estimators based on low-sampling-rate y_1 data.

The PCA+OE estimator obtained with $a = 6$ components was

$$\hat{y}_1 = \frac{\begin{array}{c}(0{\cdot}95 + 0{\cdot}44q^{-1})\hat{\tau}_1 + (2{\cdot}14 + 0{\cdot}80q^{-1})\hat{\tau}_2 \\ +(0{\cdot}50 - 0{\cdot}05q^{-1})\hat{\tau}_3 + (1{\cdot}20 + 1{\cdot}00q^{-1})\hat{\tau}_4 \\ +(1{\cdot}35 + 0{\cdot}67q^{-1})\hat{\tau}_5 + (0{\cdot}15 - 0{\cdot}88q^{-1})\hat{\tau}_6\end{array}}{1 + 0{\cdot}22q^{-1}} \times 10^{-3} \tag{25}$$

Note that although the initial pole placement was at $q = 0{\cdot}25$, the identified estimator has its pole at $z = -0{\cdot}22$. This means an oscillatory step response with a relative damping coefficient $\zeta = 0{\cdot}44$ and a natural frequency $\omega_n = 1/2T_2 = 0{\cdot}0943$ s [24].

The corresponding result for the optimal number of components $a = 4$ when the estimator was identified by use of the high-sampling-rate y_1 data was

$$\hat{y}_1 = \frac{\begin{array}{c}(0{\cdot}98 + 0{\cdot}42q^{-1})\hat{\tau}_1 + (2{\cdot}1 + 0{\cdot}8q^{-1})\hat{\tau}_2 \\ +(2{\cdot}5 + 1{\cdot}2q^{-1})\hat{\tau}_3 + (2{\cdot}3 + 0{\cdot}9q^{-1})\hat{\tau}_4\end{array}}{1 + 0{\cdot}26q^{-1}} \times 10^{-3} \tag{26}$$

The RMSEP values for the PCA+OE estimator were then 16% lower than the optimal value in the present multirate sampling case [13].

A typical PCA+OE validation response for the optimal number of components is shown in Figure 5.

Note that the estimated response follows the (in this case) measured response well also between the sampled values. This is also the case for the low-sampling-rate PLSR estimator, although the dynamic estimator gives a considerably lower validation RMSEP value.

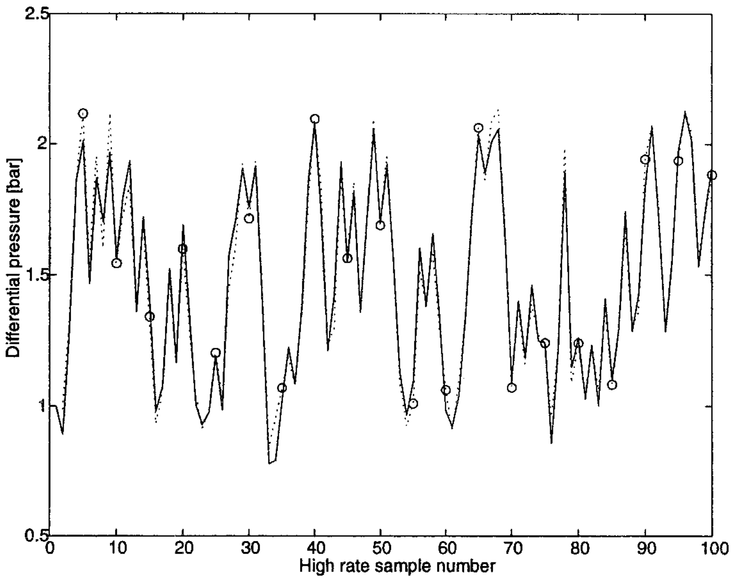

Figure 5. Validation response for dynamic PCA+OE estimator based on low-sampling-rate y_1 data. The low-sampling-rate validation data are shown as open circles, while the (in this case) known intermediate validation values are shown by the dotted curve. The estimated values are shown by the full curve.

8. CONCLUSION

High-sampling-rate $\mathbf{X}$ and low-sampling-rate $\mathbf{y}$ time series data may be used to identify an output error (OE) response estimator based on an underlying Kalman filter. The $\mathbf{X}$ variables may be both known system inputs $\mathbf{u}_k$ and secondary system outputs $\mathbf{y}_{2,k}$, while the $\mathbf{y}$ variables are the primary responses $y_{1,j}$.

A high number of collinear $\mathbf{X}$ variables may be compressed by principal component analysis (PCA), resulting in a novel PCA+OE method.

A multirate sampling example with acoustic power spectrum data demonstrates the feasibility of this method.

REFERENCES

1. Martens H, Næs T. *Multivariate Calibration*. Wiley: Chichester; 1989.
2. Ljung L. *System Identification: Theory for the User*, Prentice-Hall: Englewood Cliffs, NJ, 1999.
3. Ricker NL. The use of biased least-squares estimators for parameters in discrete-time pulse-response models. *Ind. Engng Chem. Res.* 1988; **27**: 343–350.
4. MacGregor JF, Kourti T, Kresta J. Multivariate identification: a study of several methods. *Proc. IFAC Symp. ADCHEM'91*, Toulouse, 1991; 369–375.
5. Wise BM, Ricker NL. Identification of finite response models by principal components regression: frequency-response properties. *Process Control Qual.* 1992; **4**: 77–86.
6. Dayal BS, MacGregor JF. Identification of FIR models: methods and robustness issues. *Ind. Engng Chem. Res.* 1996; **35**: 4078–4090.
7. Ergon R. On primary output estimation by use of secondary measurements as input signals in system identification. *IEEE Trans. Automatic Control* 1999; **44**: 821–825.
8. Wise BM. Adapting multivariate analysis for monitoring and modeling dynamic systems. *PhD Dissertation*, University of Washington, 1991.
9. Qin SJ, McAvoy TJ. A data-based process modeling approach and its applications. *IFAC Dynamics and Control of Chemical Reactors* (*DYCORD*+92), MD, 1992; 93–98.
10. Wise BM, Holt BR, Gallagher NB, Lee S. A comparison of neural networks, non-linear biased regression and a genetic algorithm for dynamic model identification. *Chemometrics Intell. Lab. Syst.* 1995; **30**: 81–89.
11. Dayal S, MacGregor JF. Recursive exponentially weighted PLS and its applications to adaptive control and prediction. *J. Process Control* 1997; **7**: 169–179.
12. Wikström C, Albano C, Eriksson L, Fridén H, Johansson E, Nordahl Å, Rännar S, Sandberg M, Kettaneh-Wold N, Wold S. Multivariate process and quality monitoring applied to an electrolysis process. Part II. Multivariate time-series analysis of lagged latent variables. *Chemometrics Intell. Lab. Syst.* 1998; **42**: 233–240.
13. Harnett MK, Lightbody G, Irwin GW. Dynamic inferential estimation using principal components regression (PCR). *Chemometrics Intell. Lab. Syst.* 1998; **40**: 215–224.
14. Grewal MS, Andrews AP. *Kalman Filtering: Theory and Practice*. Prentice-Hall: Englewood Cliffs, NJ, 1993.
15. Söderström T, Stoica P. *System Identification*. Prentice-Hall: Cambridge, 1989.
16. Ergon R. Dynamic system multivariate calibration by system identification methods. *Model, Ident. Control* 1998; **19**: 77–97.
17. Ergon R. Dynamic system multivariate calibration. *Chemometrics Intell. Lab. Syst.* 1998; **44**: 135–146.
18. Ergon R. Dynamic system multivariate calibration for optimal primary output estimation. *PhD Thesis*, Norwegian University of Science and Technology, Trondheim/Porsgrunn, 1999.
19. Ergon R. Dynamic system calibration: the low primary output sampling rate case. *Model. Ident. Control* 1998; **19**: 99–107.
20. Di Ruscio D. A method for identification of combined deterministic–stochastic systems. In *Applications of Computer Aided Time Series Modeling*, Aoki M, Havenner AM (eds). Springer: New York, 1997; 181–235.
21. Ljung L. *System Identification Toolbox for Use with Matlab*. Math Works: Natick, MA, 1995.
22. Johnson RA, Wichern DW. *Applied Multivariate Statistical Analysis*. Prentice-Hall: Englewood Cliffs, NJ, 1992.

23. Esbensen K, Hope B, Lied TT, Halstensen M, Gravermoen T, Sundberg K. Acoustic chemometrics for fluid flow quantifications—II: A small constriction will go a long way. *J. Chemometrics* 1999; **13**: 209–236.
24. Franklin GF, Powell JD, Workman ML. *Digital Control of Dynamic Systems*. Addison-Wesley: Reading, MA, 1990.