Estimation, System Identification and Chemometrics

Rolf Ergon, David Di Ruscio, Kim Esbensen and Svein Thore Hagen

Telemark University College, Norway, http://www-pors.hit.no/~rolfe/
Telephone: ++ 47 35 57 51 60, fax: ++ 47 35 57 52 50, e-mail address: Rolf.Ergon@hit.no

Abstract: In our master degree program in process automation, traditional modeling and control courses are supplemented by courses in experimental design and chemometrics. A corresponding inter-disciplinary research program supports this innovation in curricular structure. The background for all this is the recent developments in chemometrics and the strong stand this discipline has in the Scandinavian process engineering and industrial communities. The program curriculum and related research results are presented, together with a summary of student and industrial feedback.

Keywords: Estimation, system identification, chemometrics

1. Introduction

The M.Sc. degree in process automation at Telemark University College is based on a two years program on top of an undergraduate degree in automation, electronics, mechatronics or electrical power engineering. Due to the recent developments in chemometrics and the strong stand this discipline has acquired in the Scandinavian process engineering and industrial communities, the traditional modeling and control courses are supplemented by courses in experimental design and chemometrics. A corresponding inter-disciplinary research program supports this innovation in curricular structure. In the present paper we will first give a short outline of the program, and then focus on subjects in the areas of estimation, system identification and chemometrics with experimental design (ESIC). Our research activities in these areas will also be presented.

The total of 40 Norwegian credits in our M.Sc. program are at present represented by the courses in Table 1.

Table 1. Courses in Master degree program in Process Automation (^{o} = optional)

Applied numerical analysis	2	Structures of industrial control systems	2^o
Process control	2	*Advanced chemometrics*	2^o
Process modeling I (mechanistic)	3	Advanced control topics	2^o
Modern sensors in systems	2	Case studies in sensors and systems	2^o
Process data technology	3	Operational reliability and safety	2
State and parameter estimation	2	Project administration	1
Process modeling II (mechanistic)	2	*Project* (group assignment)	3
Chemometrics and experimental design	2	Technology in society	1
System identification and predictive control	2	*Master thesis*	9

In the table we have italicized the subjects in the ESIC area: 5 compulsory credits, 2 optional credits, 3 credits of project assignment (more or less ESIC) and 9 credits of master thesis assignment (more or less ESIC). The interested student might thus use close to one full year studying these subjects.

The central theme in the *State and parameter estimation* course is Kalman filtering. This is applied on problems concerning state estimation, parameter estimation by use of augmented and extended Kalman filters, identification of ARX models by use of recursive least squares (LS) Kalman filter algorithms, and identification of ARMAX models by use of innovations representation and the prediction error method. The essential subject of persistent excitation and its relation to statistical experimental design is also discussed. Our recent research results in this area are related to the identification of product quality estimators based on secondary plant measurements [1-7]. This also includes the handling of highly multivariate data by use of chemometrical methods.

A main subject in the *Chemometrics with experimental design* course is multivariate calibration by use of partial least squares regression (PLSR). As a basis for this, standard methods for statistical experimental design are included. Special process related applications and advanced PLSR methods are presented in the optional *Advanced chemometrics* course. Our research in the area is focused on utilization of acoustic plant information [8-10].

The *System identification* course focuses on the modern subspace identification methods, that just as the chemometrical methods utilizes projection of multivariate data onto various subspaces. This is related to chemometrics by the use of PLSR as a factorization method, and by a discussion of the important plant excitation issue. Our recent research results are related to system identification as such and to the PLSR algorithm [11-14].

It is an essential part of our program that the responsibilities of teaching the estimation and system identification courses are combined with active research on the relations between these classical control subjects and chemometrics, including industrial applications. The ESIC subjects presented above are also to various degrees applied in project and master thesis assignments based on problems from industrial partners as Norsk Hydro, Borealis, Norske Skog and Norcem, that are all major Norwegian process industry companies. A typical example is presented below.

2. Multivariate calibration

2.1 The basic problem

The central problem in the compulsory chemometrics course is the static multivariate calibration problem [15,16]. Assuming a static system with a scalar primary output or response variable y_1 and multivariate secondary y_2 outputs, the calibration problem is to find an estimator $\hat{b}$ from experimental data that may be used to estimate non-measured primary outputs according to $\hat{y}_1 = y_2^T \hat{b}$.

A typical example is the estimation of protein content in whole wheat kernels based on near infrared (NIR) spectroscopy [17]. Here, the protein content is the primary output y_1, while the NIR reflectance at a large number of frequencies gives rise to the y_2 variables. A process related example is the estimation of distillation product composition from a number of temperature measurements along the distillation tower [18]. The fundamental problem in such cases is that the number of y_2 variables may be much larger then the number of observations in the experimental data.

Assuming experimental data from independent observations, $y_1 = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1N} \end{bmatrix}^T$ and $Y_2 = \begin{bmatrix} y_{21} & y_{22} & \cdots & y_{2N} \end{bmatrix}^T$, and independent observation errors, we find the LS solution

$$\hat{b}_{LS} = \left(Y_2^T Y_2\right)^{-1} Y_2^T y_1. \tag{1}$$

With a large number of y_2 variables, this solution will be very noise sensitive, an in practical applications the LS method will work satisfactorily only when the number of variables is much smaller than the number of observations.

2.2 The chemometrical solutions

In many practical situations, fortunately, the y_2 variables are highly collinear, and the information in a large number of y_2 variables may then be compressed into a much smaller number a of estimated latent variables $\hat{\boldsymbol{t}} = \begin{bmatrix} \hat{\boldsymbol{t}}_1 & \hat{\boldsymbol{t}}_2 & \cdots & \hat{\boldsymbol{t}}_a \end{bmatrix}^T$. The model underlying such data compression is the latent variables (LV) model

$$\begin{aligned} y_1 &= L_1 \boldsymbol{t} + e_1 \\ y_2 &= L_2 \boldsymbol{t} + e_2, \end{aligned} \tag{2}$$

where e_1 and e_2 are independent observation errors.

In the chemometrical PLSR and principal component regression (PCR) methods, the Y_2 data matrix is compressed into a score matrix $\hat{T}$ by use of the factorization

$$Y_2 = \hat{T}\hat{W}_a^T + E\,, \tag{3}$$

where E is a residual matrix. Here $\hat{W}_a^T\hat{W}_a = I_a$, where a is the number of principal components one decides to use, and the least squares solution of (3) is thus $\hat{T} = Y_2\hat{W}_a$. The Y_2 data is thus projected onto a low dimensional subspace defined by $\hat{W}_a$, and the data compression results in the regularized latent variables estimator [2,19]

$$\hat{b}_{LV} = \hat{W}_a\left(\hat{W}_a^T Y_2^T Y_2 \hat{W}_a\right)^{-1}\hat{W}_a^T Y_2 y_1. \tag{4}$$

The alternative regularization method ridge regression [20] may give quite similar end results. The advantage with the chemometrical methods is, however, the interpretability of the latent variables involved [21], and this is an important part of our compulsory chemometrics course.

2.3 Optimal regularization

The static model (2) may after a similarity transformation be represented by the dynamic model

$$\begin{aligned} x_{k+1} &= v_k \\ y_{1,k} &= C_1 x_k + e_{1,k} \\ y_{2,k} &= C_2 x_k + e_{2,k}\,, \end{aligned} \tag{5}$$

where v_k, $e_{1,k}$ and $e_{2,k}$ are white noise sequences with covariances R_v, r_{11} and R_{22}.

Standard Kalman filtering theory [22] then results in the optimal estimator

$$\hat{b}_{KF} = K^T\left(KY_2^T Y_2 K^T\right)^{-1} KY_2^T y_1\,, \tag{6}$$

where

$$K = R_v C_2^T\left(C_2 R_v C_2^T + R_{22}\right)^{-1}. \tag{7}$$

The optimal weighting matrix in (4) is thus $\hat{W}_a = K^T Q$, where a is equal to the number of state variables in (5), and where Q is an invertible matrix. However, an implementation of the optimal estimator would require a detailed knowledge of the data generating system, including the process and measurement noise covariances, which especially in multivariate cases may be quite unrealistic. In practice we must thus be content with $\hat{W}_a \approx K^T Q$.

After the singular value decomposition $K = USV^T = U\begin{bmatrix} S_1 & 0\end{bmatrix}\cdot\begin{bmatrix} V_1 & V_2\end{bmatrix}^T = US_1V_1^T$, and since US_1 is invertible, we find that the optimal estimator (6) may be written as

$$\hat{b}_{KF} = V_1\left(V_1^T Y_2^T Y_2 V_1\right)V_1^T Y_2^T y_1. \tag{8}$$

Since $V_1^T V_1 = I_a$ this estimator is quite similar to (4), and it is in fact possible to show that the columns of $\hat{W}_a$ are rotated and noise corrupted versions of the V_1 columns [7].

2.4 Non-iterative PLSR algorithm

The well established PLSR algorithm is iterative in the sense that $\hat{W}_a$ is computed from the data column by column [15]. It can be shown, however, that the controllability (Krylov) matrix

$$K_a = \begin{bmatrix} Y_2^T y_1 & \left(Y_2^T Y_2\right)Y_2^T y_1 & \cdots & \left(Y_2^T Y_2\right)^{a-1} Y_2^T y_1 \end{bmatrix} \tag{9}$$

may replace $\hat{W}_a$ in (4), resulting in a non-iterative PLSR algorithm [13,14]. This may furthermore be extended to a novel non-iterative and optimal PLSR algorithm that incorporates multivariate y_1 data [13,14].

3. Dynamic system multivariate calibration

The dynamic model (5) is a special case of the more general dynamic model

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + Gv_k \\ y_{1,k} &= C_1 x_k + D_1 u_k + w_{1,k} \\ y_{2,k} &= C_2 x_k + D_2 u_k + w_{2,k}\,, \end{aligned} \tag{10}$$

where u_k are known system inputs, while v_k, $w_{1,k}$ and $w_{2,k}$ are white noise sequences. A Kalman filter with u_k and $y_{2,k}$ used as inputs, will in this dynamic case result in the optimal primary output estimate

$$\hat{y}_{1,k|k} = C_1(I - KC_2)(qI - A + AKC_2)^{-1}\left[(B - AKD_2)u_k + AKy_{2,k}\right] + C_1K(y_{2,k} - D_2u_k) + D_1u_k\,, \tag{11}$$

where q^{-1} is the unit time delay operator, and where the subscript in $\hat{y}_{1,k|k}$ indicates that data up to and including time step k is used. The optimal estimator (11) may be identified from sampled data by use of a prediction error method [23,1], and this may be done also in the many practical multirate sampling cases where the experimental data has y_1 values sampled only at a low and possibly irregular rate [3,6]. The requirement is that u and y_2 are sampled at a high enough rate to capture the dynamics of the system.

In cases where y_2 is multivariate and collinear, y_2 in (11) may be replaced by $\hat{t} = \hat{W}^T y_2$ [6].

4. Subspace identification

Subspace system identification (4SID) algorithms make use of projections of experimental data onto low dimensional subspaces, and thus have an important feature in common with the chemometrical multivariate calibration methods for static systems. The basic problem is in this case to identify the dynamic system (10) as such, with a common y_k output.

In the best-known 4SID methods [24], the first step is to identify the extended observability matrix

$$O_r = \begin{bmatrix} C^T & (CA)^T & \cdots & (CA^{r-1})^T \end{bmatrix}^T, \tag{12}$$

and after the appropriate projections this can be done by LS methods. When an estimate $\hat{O}_r$ is determined, $\hat{C}$ is obtained as the first block row, while $\hat{A}$ easily follows form the shift property of $\hat{O}_r$, i.e. the fact that $\begin{bmatrix} O_{r-1}^T & (CA^{r-1})^T \end{bmatrix}^T = \begin{bmatrix} C^T & (O_{r-1}A)^T \end{bmatrix}^T$. With $\hat{A}$ and $\hat{C}$ in place, it is an easy LS problem to find $\hat{B}$ and $\hat{D}$ from

$$y_k = \hat{C}(qI - \hat{A})^{-1} Bu_k + Du_k + w_k\,. \tag{12}$$

By use of $\hat{O}_r$ it is also possible to reconstruct the states x_k in (10), and the statistical noise properties can then be established.

In the DSR algorithm of Di Ruscio [11,12] the first step is to eliminate x_k from the equations, and also here the shift property of O_r plays a vital role. After appropriate projections, estimates $\hat{A}$, $\hat{B}$, $\hat{C}$ and $\hat{D}$ may then be found by LS methods, although QR factorization plays a vital role in the practical implementation of the algorithm.

The noise properties, i.e. the innovations covariance Λ and the gain K in an underlying Kalman filter, are then also estimated.

5. Master thesis example

The quality control of polymer production processes is a quite active research area [25]. In a present master thesis assignment related to the Borealis polyolefine plant in Bamble, Norway, the problem is to identify melting index estimators from plant data. The task includes data reconciliation, static PLSR modeling and validation, dynamic multirate estimator identification, and discussions of estimator updating and feedback control schemes.

6. Course program evaluation

The course program in Table 1 is evaluated each semester. This has the form of discussions in the classes, followed by a formal meeting between student representatives and all teaching personnel involved. Our experiences with this form of oral evaluation are quite good, and the student response to the program is overall positive. The response in the industrial community is also quite favorable, as illustrated by the fact that around 90% of all master thesis assignments are given in close cooperation with industrial partners. Many of those are in the ESIC area. Post-graduation feedback also indicates that the inclusion of experimental design and chemometrics in the curriculum is well motivated.

References

[1] R. Ergon and D. Di Ruscio, "Dynamic system calibration by system identification methods", Proc. Fourth European Control Conference (EEC'97), Brussels, Belgium, CD-ROM, 1997

[2] R. Ergon , "Dynamic system multivariate calibration by system identification methods", Modeling, Identification and Control, Vol. 19, No. 2, pp 77-97, 1998

[3] R. Ergon, "Dynamic system calibration: The low primary output sampling rate case", Modeling, Identification and Control, Vol. 19, No. 2, pp. 99-107, 1998

[4] R. Ergon, "Dynamic system multivariate calibration", Chemometrics and Intelligent Laboratory Systems, Vol. 44, pp. 135-146, 1998

[5] R. Ergon, "On primary output estimation by use of secondary measurements as input signals in system identification", IEEE Transactions on Automatic Control, Vol. 44, No. 4, pp. 821-825, 1999

[6] R. Ergon, "Dynamic System Multivariate Calibration for Optimal Primary Output Estimation", Ph.D. thesis, The Norwegian University of Science and Technology/Telemark University College, Trondheim/Porsgrunn, Norway, 1999

[7] R. Ergon, "Multivariate calibration in a Kalman filtering perspective", submitted to Journal of Chemometrics, 2000

[8] K. Esbensen, B. Hope, T.T. Lied, M. Halstensen, T. Gravermoen and K. Sundberg, "Acoustic chemometrics for fluid flow quantifications-II: A small constriction will go a long way", Journal of Chemometrics, 13, pp. 1-29, 1999

[9] M. Halstensen and K. Esbensen, "New developments in acoustic chemometric prediction of particle size distribution - 'The problem is the solution'", accepted for publication in Journal of Chemometrics, 2000

[10] R. Ergon and M. Halstensen, "Dynamic system multivariate calibration with low sampling rate y data", accepted for publication in Journal of Chemometrics, 2000

[11] D. Di Ruscio, "A method for identification of combined deterministic-stochastic systems", in "Applications of Computer Aided Time Series Modeling", M. Aoki and A.M. Havenner, Eds., Springer- Verlag, New York, 1997

[12] D. Di Ruscio, "On Subspace Identification of the Extended Observability Matrix", Proc. of the 36th IEEE CDC, San Diego, 1997

[13] D. Di Ruscio, "The partial least squares algorithm: A truncated Cayley-Hamilton series approximation used to solve the regression problem", Modeling, Identification and Control, Vol. 19, No. 3, pp. 117-1408, 1998

[14] D. Di Ruscio, "A weighted view on the partial least-squares algorithm", Automatica, Vol. 36, pp. 831-850, 2000

[15] H. Martens and T. Næs, "Multivariate Calibration", John Wiley & Sons, New York, 1989

[16] K. Esbensen, "Multivariate Data Analysis - in practice", Camo ASA, Trondheim, Norway, 2000

[17] K.H. Norris, "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, 1993

[18] T. Mejdell and S. Skogestad, "Estimate of process output from multiple secondary measurements", Proc. American Control Conference, 2112-2121, 1989

[19] D. Di Ruscio, "Subspace System Identification: Theory and applications", Lecture notes, Telemark University College, Porsgrunn, Norway, 1997

[20] A.E. Hoerl and R.W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems", Technometrics, Vol. 12, No. 1, Pp. 55-67, 1970

[21] S. World, "Discussion: PLSR in chemical practice", Techno metrics, Vol. 35, No. 2, pp. 136-139, 1993

[22] M.S. Growl and ASP. Andrews, "Kalman Filtering: Theory and Practice", Prentice Hall, New Jersey, 1993

[23] L. Lung, "System Identification: Theory for the user", Prentice-Hall, New Jersey, 1999

[24] P. Van Overshoe and B. De Moor, "Subspace Identification for Linear Systems", Lower Academic Publishers, Correct, The Netherlands, 1996

[25] M. Ohshima and M. Tanoak, "Quality control of polymer production processes", Journal of Process Control, Vol. 10, pp. 135-148, 2000