A new class of model-based static estimators

Maryam Ghadrdan[1], Chriss Grimholt[1], Sigurd Skogestad[1]

1 Department of Chemical Engineering, Norwegian University of Science and Technology, N-7491 Trondheim, Norway, Email: ghadrdan@nt.ntnu.no, grimholt@stud.ntnu.no, skoge@nt.ntnu.no

Abstract

Static estimators are commonly used as "soft-sensors" in the process industry. The performance of the estimators depend on whether it is used for monitoring (open-loop) or for closed-loop control applications. In this work, we propose to design the estimators which are specialized for each case. The approach is to minimize the estimation error for expected disturbances and measurement noise. The main extension compared to previous work is to include measurement noise and to provide explicit formulae for computing the optimal static estimator. We also compare the results with standard existing estimators, e.g. PLS. The approach is applied to estimation of product composition in a distillation column from combination of temperature measurements.

Keywords: estimation, static estimator, combination of measurements, distillation, composition

1. Introduction

In a chemical plant, there are usually a large number of sensors which are used for monitoring and control of processes. However, some process variables (e.g., composition) may be too difficult or expensive to measure online. Estimators, also called soft sensors, work by predicting such variables using available measurements (e.g. temperatures).

Both dynamic and static estimators may be used, but the simpler static estimators are most common in the process industry. Since our method is a static estimator, our literature survey is limited to this group. There are many approaches that have been used to obtain the static estimators, including multivariate regression [16, 27], artificial neural networks [2], support vector machine regression [26], etc.

Principle Component Regression (PCR) [15] and Partial Least Squares (PLS) [24] are two of the most used data analysis tools in chemometrics. These methods are based on projecting the solution to a lower-dimensional subspace. The literature review by Wentzell and Montoto [23] compared these two methods, covering both experimental and simulation studies. In short, the advantage of PLS is that the method obtains a small prediction error with fewer principal components than for PCR. Li and Shao [13] have mentioned large samples needed for regression and insensitivity to measurement errors as two drawbacks of these methods. Because of the popularity of these methods, we are going to compare our method with the PLS regression method.

The simplest model-based static estimator is the "inferential estimator" of Brosilow and coworkers ([10]). Let $\mathrm{u}' = \begin{bmatrix} \mathrm{u} & \mathrm{d} \end{bmatrix}$ represent the vector of independent variables, including the inputs u and the disturbances d. Let x represent the process measurements and y the variables we want to estimate. Let the linear static model in deviation variables be

$$\mathrm{x} = \mathbf{X}\,\mathrm{u}' \tag{1}$$

$$\mathrm{y} = \mathbf{Y}\,\mathrm{u}' \tag{2}$$

The "Brosilow" estimator is then simply the following least squares estimate of y

$$\hat{\mathrm{y}} = \mathbf{H}\mathrm{x}_m \tag{3}$$

where

$$\mathbf{H} = \mathbf{Y}\mathbf{X}^{\dagger} \tag{4}$$

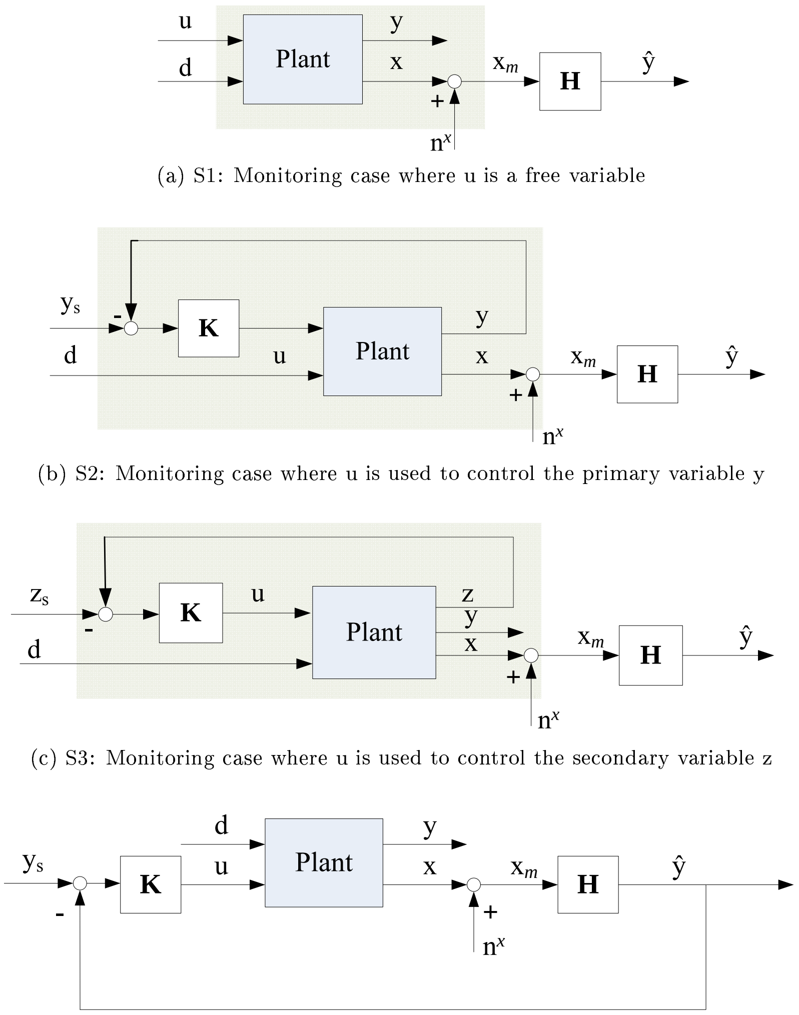
and $\mathbf{X}^{\dagger}$ is the pseudo inverse of the matrix $\mathbf{X}$.

Joseph and Brosilow [10] discuss some of the weaknesses of this estimator. For "ill-conditioned" plants, they find that the estimate may be improved in some cases by removing measurements, because this reduces the condition number. Where the condition number of $\mathbf{X}$ is large, intuitively, removing measurements cannot be the optimal way of dealing with these problems, because we are then throwing away information. This is also clear when we consider the popular "data-based" regression estimators, like Partial Least Squares (PLS) regression [12], where one does not remove measurements, but instead removes weak "directions" in the data.

A fundamental problem with the Brosilow inferential estimator is that it fails to take into account measurement noise in an explicit manner. The main goal of this paper is to include the effect of measurement noise in the derivation of the optimal model-based static estimator. This is the first static estimator to include static measurement error (noise) in a rigorous manner. This allows one to design an estimator which is optimal in terms of disturbances and also in terms of noise. The latter means that we handle in an optimal manner the "high condition number problem", which has been a major concern in previous work ([10], [16], [9], [17], [22]).The derivation is straightforward, but surprisingly it seems it has not been presented before.

Another issue is that the Brosilow least squares estimator does not take into account whether the estimator is used only for monitoring or for closed-loop operation. Actually the latter is a shortcoming of most existing data-based estimators. In the paper, we derive optimal estimators for four cases as illustrated in Figure 1. Case S1 is the direct extension of the Brosilow inferential estimator to include measurement noise. In case S2, the inputs u are used to control the variables y at given setpoint y_s. It is similar to case S1, except that the setpoint y_s takes the role of the inputs. Case S3 is a generalization where we control the variables z. Cases S1, S2 and S3 are practically relevant if the estimator is used for monitoring only, because the estimate ŷ is not used for control. Finally, case S4 is the relevant case when we use the estimator in closed loop (for control purposes). Whereas the optimal estimators for cases for cases S1, S2 and S3 are least-square estimators with a similar structure to the Brosilow estimator in (4), the structure for case S4 is quite different and the mathematics to derive it are more complex. The derivation is based on results for optimal measurement combination for self-optimizing control [1] and is the main new contribution of this paper.

The derivation of the new static estimators is presented in section 2. The concept of some well-known data-based estimators are described in section 3. In section 4 we discuss how we can use our new ideas for optimal model-based to derive new data-based estimator. Finally, in section 5, we compare the new static estimators with previous work, including the Brosilow estimator and regression based estimators on distillation case-studies.



(a) S1: Monitoring case where u is a free variable

(b) S2: Monitoring case where u is used to control the primary variable y

(c) S3: Monitoring case where u is used to control the secondary variable z

(d) S4= CL: "closed-loop" case where u is used to control the predicted variable ŷ

Figure 1: Block diagrams for different scenarios

2. Derivation of Model-based Static Estimators

2.1. Problem definition

In this section, we derive optimal "open-loop" and "closed-loop" static estimators. By "optimal", it is meant that we for a linear estimator of the form

$$\hat{y} = \mathbf{H}x_m \tag{5}$$

want to minimize the expected prediction error

$$e = y - \hat{y} \tag{6}$$

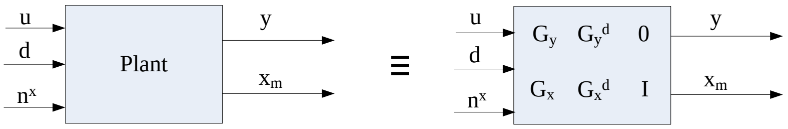


Figure 2: Block diagram for the linearized plant

The actual measurements x_m, containing measurement noise n^x are

$$\mathrm{x}_m = \mathrm{x} + \mathrm{n}^x \tag{7}$$

We define the following variables:

- u: inputs (degrees of freedom), these may represent setpoints for lower-layer controllers
- d: disturbances, including parameter changes.
- x: all available measured variables.
- n^x: measurement noise (error) for x.
- y: primary variables that we want to estimate
- z: secondary variables, which we may control, $\dim(\mathrm{z}) = \dim(\mathrm{u})$

All variables are assumed to be deviation variables (away from the nominal or centered values). We use linear static models for the primary variables y, measurements x, and secondary variables z (see Figure 2)

$$\mathrm{y} = \mathbf{G}_y \mathrm{u} + \mathbf{G}_y^d \mathrm{d} \tag{8}$$

$$\mathrm{x} = \mathbf{G}_x \mathrm{u} + \mathbf{G}_x^d \mathrm{d} \tag{9}$$

$$\mathrm{z} = \mathbf{G}_z \mathrm{u} + \mathbf{G}_z^d \mathrm{d} \tag{10}$$

In terms of the notation used for the Brosilow inferential estimator in (1) we have

$$\mathbf{X} = \left[\begin{array}{cc} \mathbf{G}_x & \mathbf{G}_x^d \end{array}\right] \tag{11}$$

$$\mathbf{Y} = \left[\begin{array}{cc} \mathbf{G}_y & \mathbf{G}_y^d \end{array}\right] \tag{12}$$

In addition, the expected magnitude variables of the independent variables for each case (see Figure 1) is quantified by weighting matrices ($\mathbf{W}_u$, $\mathbf{W}_d$, $\mathbf{W}_{n^x}$,$\mathbf{W}_{y_s}$, $\mathbf{W}_{z_s}$), as explained in detail below.

2.2. Estimators used for monitoring (cases S1, S2 and S3)

With the term “open-loop”, it is implied that the predicted variables $\hat{y} = \mathbf{H}\mathrm{x}_m$ are used for monitoring, that is, they are not used for control purposes. It should be noted that this is not the same as implying that the variables in a given system are uncontrolled. We can think of three main types of open-loop monitoring estimators are illustrated in Figure 1:

Case S1 Predicting primary variables from a system with no control, i.e. the inputs u are free variables.

Case S2 Predicting primary variables from a system where the primary variables y are measured and controlled, i.e. the inputs u are used to keep $\mathrm{y} = \mathrm{y}_s$.

Case S3 Predicting primary variables from a system where the inputs u are used to control the secondary variables z, i.e. $\mathrm{z} = \mathrm{z}_s$.

We first consider case S1 in detail. Cases S2 and S3 are then straightforward extensions.

2.2.1. Case S1

Case S1 is the direct extension of the Brosilow estimator to include noise. To find the optimal estimator for open-loop operation, the prediction error has to be expressed as a function of the system and the estimator.

Lemma 1. *For a given linear estimator , when applied to the system defined in equations (5)-(9), and considering that* u *is a free variable, the prediction error can be expressed as*

$$e\left(\mathbf{H}\right) = \left[\begin{array}{ccc} \left(\mathbf{G}_y - \mathbf{H}\mathbf{G}_x\right) & \left(\mathbf{G}_y^d - \mathbf{H}\mathbf{G}_x^d\right) & -\mathbf{H} \end{array}\right] \left[\begin{array}{c} \mathrm{u} \\ \mathrm{d} \\ \mathrm{n}^x \end{array}\right] \tag{13}$$

Proof. An expression of ŷ as a function of u, d and n^x can be obtained by substituting equations (7) and (9) into equation (5).

$$\hat{\mathrm{y}} = \mathbf{H}\left(\mathbf{G}_x\mathrm{u} + \mathbf{G}_x^d\mathrm{d} + \mathrm{n}^x\right)$$

Using the definition of prediction error and substituting the expression for ŷ, we will have

$$\mathrm{e}\left(\mathbf{H}\right) = \left(\mathbf{G}_y - \mathbf{H}\mathbf{G}_x\right)\mathrm{u} + \left(\mathbf{G}_y^d - \mathbf{H}\mathbf{G}_x^d\right)\mathrm{d} - \mathbf{H}\mathrm{n}^x$$

which is the same as equation (13). □

Next, we derive an expression for the expected prediction error, assuming that u, d, n^x are normally distributed with given weight matrices.

Lemma 2. ***Expected prediction error.*** *Let the disturbance and noise be normalized on the form*

$$\mathrm{u} = \mathbf{W}_u \mathrm{u}'$$

$$\mathrm{d} = \mathbf{W}_d \mathrm{d}'$$

$$\mathrm{n}^x = \mathbf{W}_{n^x} \mathrm{n}^{x'}$$

where the elements u', d' and $n^{x'}$ of the normalized vectors u', d' *are assumed to be normally distributed with zero mean and unit standard deviation;*

$$\begin{aligned} u' &\sim \mathcal{N}(0,1) \\ d' &\sim \mathcal{N}(0,1) \\ n^{x'} &\sim \mathcal{N}(0,1) \end{aligned}$$

The diagonal scaling matrices $\mathbf{W}_u$, $\mathbf{W}_d$ and $\mathbf{W}_{n^x}$ contain the standard deviations of the elements in u, d *and* n^x *respectively.*

From Lemma 1 the prediction error can be expressed as

$$e = \underbrace{\left[\begin{array}{ccc} (\mathbf{G}_y - \mathbf{H}\mathbf{G}_x)\,\mathbf{W}_u & \left(\mathbf{G}_y^d - \mathbf{H}\mathbf{G}_x^d\right)\mathbf{W}_d & -\mathbf{H}\mathbf{W}_{n^x} \end{array}\right]}_{\mathbf{M}(\mathbf{H})} \left[\begin{array}{c} \mathrm{u}' \\ \mathrm{d}' \\ \mathrm{n}^{x'} \end{array}\right]$$

The expected value of the 2-norm of the prediction error (variance) then becomes

$$\mathrm{E}\left(\|e\|_2\right) = \|\mathbf{M}(\mathbf{H})\|_F^2$$

Proof. Let $\tilde{\mathrm{d}} = \left[\begin{array}{c} \mathrm{u}' \\ \mathrm{d}' \\ \mathrm{n}^{x'} \end{array}\right]$. Then, $\mathrm{e} = \mathbf{M}\tilde{\mathrm{d}}$, and noting that $\|\mathrm{e}\|_2 = \mathrm{tr}\left(\mathrm{e}\mathrm{e}^T\right)$, the expected value of the 2-norm of the prediction error can be written as

$$\begin{aligned} \mathrm{E}\left(\|\mathrm{e}\|_2\right) &= \mathrm{E}\left[\mathrm{tr}\left(\mathbf{M}\tilde{\mathrm{d}}\tilde{\mathrm{d}}^T\mathbf{M}^T\right)\right] \\ &= \mathrm{E}\left[\mathrm{tr}\left(\mathbf{M}^T\mathbf{M}\tilde{\mathrm{d}}\tilde{\mathrm{d}}^T\right)\right] \\ &= \mathrm{tr}\left(\mathbf{M}^T\mathbf{M}\,\mathrm{E}\left[\tilde{\mathrm{d}}\tilde{\mathrm{d}}^T\right]\right) \end{aligned}$$

where $\mathrm{tr}(.)$ denotes the trace of the matrix and $E[.]$ is the expectation operator.

Since $\left\|\left[\begin{array}{c} \mathrm{u}' \\ \mathrm{d}' \\ \mathrm{n}^{x'} \end{array}\right]\right\| \sim \mathcal{N}\left(0, \mathbf{I}_{n_u+n_d+n_x}\right)$, by substituting the normal distribution in the definition of expected value we have

$$\mathrm{E}\left[\tilde{\mathrm{d}}\tilde{\mathrm{d}}^T\right] = \mathrm{Var}\left(\tilde{\mathrm{d}}\right)$$

In addition, we know that the square root of the trace of the matrix $\mathbf{M}^T\mathbf{M}$ is actually the definition of Frobenius norm of matrix $\mathbf{M}$. So,

$$\mathrm{E}\left(\|\mathrm{e}\|_2\right) = \mathrm{tr}\left(\mathbf{M}^T\mathbf{M}\right) = \|\mathbf{M}\|_F^2$$

□

From Lemma 2, the expected value of the 2-norm prediction error (variance) is minimized by selecting $\mathbf{H}$ to minimize $\|\mathbf{M}\|_F$. This leads to the following theorem

Theorem 1. *The optimal "open-loop" estimator following the linear relationship*

$$\hat{\mathrm{y}} = \mathbf{H}\mathrm{x}_m$$

that minimizes the variance of the prediction error (Lemma 1 and 2)

$$\mathrm{e} = \mathrm{y} - \hat{\mathrm{y}}$$

when u is a free variable, is

$$\mathbf{H}_1 = \mathbf{Y}_1\mathbf{X}_1^\dagger \tag{14}$$

where $\mathbf{X}^\dagger$ is the pseudo-inverse of $\mathbf{X}$, and

$$\begin{aligned} \mathbf{Y}_1 &= \begin{bmatrix} \mathbf{G}_y\mathbf{W}_u & \mathbf{G}_y^d\mathbf{W}_d & 0 \end{bmatrix} \\ \mathbf{X}_1 &= \begin{bmatrix} \mathbf{G}_x\mathbf{W}_u & \mathbf{G}_x^d\mathbf{W}_d & \mathbf{W}_{n^x} \end{bmatrix} \end{aligned} \tag{15}$$

If $\mathbf{X}_1$ has full column rank, we have $\mathbf{X}_1^\dagger = \left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}\mathbf{X}_1^T$. If $\mathbf{X}_1$ has full row rank, we have $\mathbf{X}_1^\dagger = \mathbf{X}_1^T\left(\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}$. For the general case, where $\mathbf{X}_1$ has neither full row nor column rank, the pseudo-inverse may be obtained using the singular value decomposition

Proof. In Lemma 2, we showed that minimizing $\|\mathrm{e}\left(\mathbf{H}\right)\|_2$ is equivalent to minimizing $\|\mathbf{M}\left(\mathbf{H}\right)\|_F^2$ for the expected prediction error. $\mathbf{M}(\mathbf{H})$ can be rewritten as

$$\mathbf{M} = \mathbf{Y}_1 - \mathbf{H}\mathbf{X}_1$$

The optimization problem then becomes

$$\min_{\mathbf{H}}\|\mathbf{Y}_1 - \mathbf{H}\mathbf{X}_1\|$$

and we recognize that this is the least squares problem with the known optimal solution.

$$\mathbf{H}_1 = \mathbf{Y}_1\mathbf{X}_1^\dagger$$

□

Figure 3 shows an interpretation of Theorem 1, which is a direct generalization of the Brosilow estimator, when we also include noise. Note that the elements in $\mathbf{Y}_1$ corresponding to $\mathrm{n}^{x'}$ is zero.

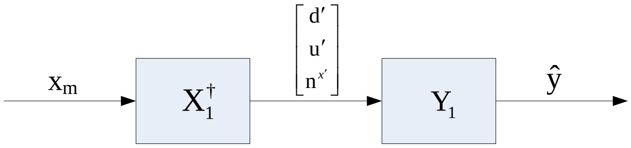

$\mathrm{x_m}$
$\mathrm{X}_1^\dagger$
$\begin{bmatrix} \mathrm{d}' \\ \mathrm{u}' \\ \mathrm{n}^{x'} \end{bmatrix}$
Y_1
$\hat{\mathrm{y}}$

Figure 3: Interpretation of Theorem 1

This estimator is optimal for the case where the process input u are truly independent variables, that is, when we have no control (case S1 in Figure 1).

2.2.2. Case S2

We now consider the case where the inputs u are used to keep the outputs y at given setpoints y_s. This means that y_s replaces u as independent variables. It is assumed that $\dim(\mathrm{y}) = \dim(\mathrm{u})$.

Theorem 2. *Optimal "open-loop" estimator for closed-loop operation (controlled* y*).*

The optimal "open-loop" estimator $\mathbf{H}$ *for closed-loop operation where the degrees of freedom* u *are adjusted such that the primary variables* y *are kept at the setpoints* y_s

$$\mathrm{y} = \mathrm{y}_s$$

when applied to the system defined above and considering the expected prediction error, is

$$\mathbf{H}_2 = \mathbf{Y}_2\mathbf{X}_2^\dagger$$

where

$$\begin{aligned} \mathbf{Y}_2 &= \begin{bmatrix} \mathbf{W}_{y_s} & 0 & 0 \end{bmatrix} \\ \mathbf{X}_2 &= \begin{bmatrix} \mathbf{G}_x^{cl}\mathbf{W}_{y_s} & \mathbf{F}\mathbf{W}_d & \mathbf{W}_{n^x} \end{bmatrix} \end{aligned} \tag{16}$$

where $\mathbf{G}_x^{cl} = \mathbf{G}_x\mathbf{G}_y^{-1}$ *and* $\mathbf{F} = \mathbf{G}_x^d - \mathbf{G}_x\mathbf{G}_y^{-1}\mathbf{G}_y^d$

Proof. Considering that u is used for keeping $\mathrm{y} = \mathrm{y}_s$. Solving equation (8) with respect to u when $\mathrm{y} = \mathrm{y}_s$ gives

$$\mathrm{u} = \mathbf{G}_y^{-1}\mathrm{y}_s - \mathbf{G}_y^{-1}\mathbf{G}_y^d\mathrm{d}$$

By combining equations (9), (7) and (5) with the above equation, the following expression for $\hat{\mathrm{y}}$ as an explicit function of y_s, d and n^x is obtained.

$$\hat{\mathrm{y}} = \mathbf{H}\left[\mathbf{G}_x\mathbf{G}_y^{-1}\mathrm{y}_s + \left(\mathbf{G}_x^d - \mathbf{G}_x\mathbf{G}_y^{-1}\mathbf{G}_y^d\right)\mathrm{d} + \mathrm{n}^x\right]$$

Here, $\left(\mathbf{G}_x^d - \mathbf{G}_x\mathbf{G}_y^{-1}\mathbf{G}_y^d\right) = \left(\frac{\partial \mathrm{x}}{\partial \mathrm{d}}\right)_{\mathrm{y}=\mathrm{y_s}}$ is the optimal sensitivity $\mathbf{F}$ [1], and $\mathbf{G}_x\mathbf{G}_y^{-1} = \left(\frac{\partial \mathrm{x}}{\partial \mathrm{y_s}}\right)_{\mathrm{d}}$ known as the closed-loop gain $\mathbf{G}_x^{cl}$. So, the above equation becomes

$$\hat{\mathrm{y}} = \mathbf{H}\left[\mathbf{G}_x^{cl}\mathrm{y}_s + \mathbf{F}\mathrm{d} + \mathrm{n}^x\right]$$

With the assumption that $\mathrm{y} = \mathrm{y}_s$, the prediction error then becomes

$$\mathrm{e} = \mathrm{y} - \hat{\mathrm{y}} = \begin{bmatrix} \left(\mathbf{I} - \mathbf{H}\mathbf{G}_x^{cl}\right) & -\mathbf{HF} & -\mathbf{H} \end{bmatrix} \begin{bmatrix} \mathrm{y}_s \\ \mathrm{d} \\ \mathrm{n}^x \end{bmatrix}$$

Proceeding analogous to Lemmas 1 and 2 and Theorem 1, results in the given preposition. □

2.2.3. Case S3

Theorem 3. *(Generalize theorems 1 and 2) Optimal "open-loop" estimator for closed-loop operation (controlled* z*). The optimal "open-loop" estimator* $\mathbf{H}$ *for closed-loop operation where the degrees of freedom* u *are adjusted such that the secondary variables* z *are kept at the setpoints* z_s

$$\mathrm{z} = \mathrm{z}_s$$

when applied to the system defined above, and considering the expected prediction error, is

$$\mathbf{H}_3 = \mathbf{Y}_3\mathbf{X}_3^{\dagger}$$

where

$$\mathbf{Y}_3 = \begin{bmatrix} \mathbf{G}_y^{cl}\mathbf{W}_{z_s} & \mathbf{F}_y'\mathbf{W}_d & 0 \end{bmatrix}$$

$$\mathbf{X}_3 = \begin{bmatrix} \mathbf{G}_x^{cl}\mathbf{W}_{z_s} & \mathbf{F}_x'\mathbf{W}_d & \mathbf{W}_{n^x} \end{bmatrix}$$

where $\mathbf{G}_y^{cl} = \mathbf{G}_y\mathbf{G}_z^{-1}$, $\mathbf{G}_x^{cl} = \mathbf{G}_x\mathbf{G}_z^{-1}$, $\mathbf{F}_y' = \mathbf{G}_y^d - \mathbf{G}_y\mathbf{G}_z^{-1}\mathbf{G}_z^d$ *and* $\mathbf{F}_x' = \mathbf{G}_x^d - \mathbf{G}_x\mathbf{G}_z^{-1}\mathbf{G}_z^d$

Proof. Considering that u is used for keeping $\mathrm{z} = \mathrm{z}_s$. Solving equation 10 with respect to u when $\mathrm{z} = \mathrm{z}_s$ gives

$$\mathrm{u} = \mathbf{G}_z^{-1}\mathrm{z}_s - \mathbf{G}_z^{-1}\mathbf{G}_z^d\mathrm{d}$$

By combining equations 8 and the above expression for u, we will have

$$\mathrm{y} = \underbrace{\mathbf{G}_y\mathbf{G}_z^{-1}}_{G_y^{cl}}\mathrm{z}_s + \left(\underbrace{\mathbf{G}_y^d - \mathbf{G}_y\mathbf{G}_z^{-1}\mathbf{G}_z^d}_{F_y'}\right)\mathrm{d}$$

Introducing the optimal sensitivity $\mathbf{F'}_y$ and the closed-loop gain $\mathbf{G}_y^{cl}$ we will get

$$\mathrm{y} = \mathbf{G}_y^{cl}\mathrm{z}_s + \mathbf{F'}_y\mathrm{d}$$

By combining equations 9, 7 and 5, the following expression for $\hat{\mathrm{y}}$ as an explicit function of y_s, d and n^x is obtained.

$$\hat{\mathrm{y}} = \mathbf{H}\left[\underbrace{\mathbf{G}_x\mathbf{G}_z^{-1}}_{G_x^{cl}}\mathrm{z}_s + \left(\underbrace{\mathbf{G}_x^d - \mathbf{G}_x\mathbf{G}_z^{-1}\mathbf{G}_z^d}_{F_x'}\right)\mathrm{d} + \mathrm{n}^x\right]$$

Using the definition of prediction error with the expression for $\hat{\mathrm{y}}$ and y gives

$$\mathrm{e}\,(\mathrm{H}) = \left[\begin{array}{ccc} \left(\mathbf{G}_y^{cl} - \mathbf{H}\mathbf{G}_x^{cl}\right) & \left(\mathbf{F'}_y - \mathbf{H}\mathbf{F}'_x\right) & -\mathbf{H} \end{array}\right]\left[\begin{array}{c} \mathrm{z}_s \\ \mathrm{d} \\ \mathrm{n}^x \end{array}\right]$$

Proceeding analogous to Lemma 2 and Theorem 1, will result in the given proposition. □

Note that Theorem 3 is a generalization of Theorems 1 and 2, since setting $\mathrm{z} = \mathrm{u}$ gives Theorem 1 and setting $\mathrm{z} = \mathrm{y}$ gives Theorem 2.

2.3. The "closed-loop" estimator

In this section, we will find an expression for the prediction error under the assumption that the prediction $\hat{\mathrm{y}} = \mathbf{H}\mathrm{x}_m$ is used for controlling the primary variables, that is, we have $\hat{\mathrm{y}} = \mathrm{y}_s$. It is assumed that $\dim(\mathrm{y}) = \dim(\mathrm{u})$.

Theorem 4. *The optimal "closed-loop" estimator $\mathbf{H}_{CL}$ following the linear relationship*

$$\hat{\mathrm{y}} = \mathbf{H}_{CL}\mathrm{x}_m \tag{17}$$

that minimizes the expected prediction error

$$e = \mathrm{y} - \hat{\mathrm{y}}$$

for the expected sets of d,n^x *and* y_s*, assuming that the degrees of freedom* u *are adjusted to keep the prediction at the setpoint ($\hat{\mathrm{y}} = \mathrm{y}_s$), is*

$$\mathbf{H}_{CL} = \arg(\min_{\mathbf{H}} \left\|\mathbf{H}\left[\begin{array}{cc} \mathbf{F}\mathbf{W}_d & \mathbf{W}_{n^x} \end{array}\right]\right\|_F) \tag{18}$$

$$s.t.\ \mathbf{H}\mathbf{G}_x = \mathbf{G}_y$$

or equivalently $\mathbf{H}\mathbf{G}_x^c l = \mathbf{I}$ *where the sensitivity matrix* $\mathbf{F}$ *is defined as*

$$\mathbf{F} = \left(\frac{\partial \mathrm{x}}{\partial \mathrm{d}}\right)_{\mathrm{y}=0} = \mathbf{G}_x^d - \mathbf{G}_x\mathbf{G}_y^{-1}\mathbf{G}_y^d$$

Proof. An expression for the prediction as an explicit function of u , d and n^x is achieved by combining Equations (9), (7) and (5) and becomes

$$\hat{\mathrm{y}} = \mathbf{H}\left(\mathbf{G}_x\mathrm{u} + \mathbf{G}_x^d\mathrm{d} + \mathrm{n}^x\right) \tag{19}$$

Assume that the predictions $\hat{\mathrm{y}}$ are held at the setpoints y_s by manipulating u. Solving equation (19) with respect to u when $\hat{\mathrm{y}} = \mathrm{y}_s$, gives

$$\mathrm{u} = -\left(\mathbf{H}\mathbf{G}_x\right)^{-1}\mathbf{H}\left(\mathbf{G}_x^d\mathrm{d} + \mathrm{n}^x\right) + \left(\mathbf{H}\mathbf{G}_x\right)^{-1}\mathrm{y}_s \tag{20}$$

An expression for the primary variable y, when u is used for keeping $\hat{\mathrm{y}} = \mathrm{y}_s$, is achieved by inserting Equation (20) into (8) and yields

$$\begin{aligned} \mathrm{y} &= -\mathbf{G}_y\left(\mathbf{H}\mathbf{G}_x\right)^{-1}\mathbf{H}\left(\mathbf{G}_x^d\mathrm{d} + \mathrm{n}^x\right) + \mathbf{G}_y\left(\mathbf{H}\mathbf{G}_x\right)^{-1}\mathrm{y}_s + \mathbf{G}_y^d\mathrm{d} \\ &= -\mathbf{G}_y\left(\mathbf{H}\mathbf{G}_x\right)^{-1}\mathbf{H}\left[\left(\mathbf{G}_x^d - \mathbf{G}_x\mathbf{G}_y^{-1}\mathbf{G}_y^d\right)\mathrm{d} + \mathrm{n}^x\right] + \mathbf{G}_y\left(\mathbf{H}\mathbf{G}_x\right)^{-1}\mathrm{y}_s \end{aligned} \tag{21}$$

where $\left(\mathbf{G}_x^d - \mathbf{G}_x\mathbf{G}_y^{-1}\mathbf{G}_y^d\right)$ is the optimal sensitivity $\mathbf{F}$, and the expression becomes

$$\mathrm{y} = -\mathbf{G}_y\left(\mathbf{H}\mathbf{G}_x\right)^{-1}\mathbf{H}\left[\mathbf{F}\mathrm{d} + \mathrm{n}^x\right] + \mathbf{G}_y\left(\mathbf{H}\mathbf{G}_x\right)^{-1}\mathrm{y}_s \tag{22}$$

Inserting the expression for y into the prediction error e, remembering that the prediction is kept at the setpoint ($\hat{\mathrm{y}} = \mathrm{y}_s$), gives

$$\begin{aligned} \mathrm{e} = \mathrm{y} - \hat{\mathrm{y}} = \mathrm{y} - \mathrm{y}_s &= -\mathbf{G}_y\left(\mathbf{H}\mathbf{G}_x\right)^{-1}\mathbf{H}\left(\mathbf{F}\mathrm{d} + \mathrm{n}^x\right) + \mathbf{G}_y\left(\mathbf{H}\mathbf{G}_x\right)^{-1}\mathrm{y}_s - \mathrm{y}_s \\ &= -\mathbf{G}_y\left(\mathbf{H}\mathbf{G}_x\right)^{-1}\mathbf{H}\left(\mathbf{F}\mathrm{d} + \mathrm{n}^x\right) + \left[\mathbf{G}_y\left(\mathbf{H}\mathbf{G}_x\right)^{-1} - \mathbf{I}\right]\mathrm{y}_s \end{aligned} \tag{23}$$

Introducing normalized variables,

$$\mathrm{e} = \underbrace{-\mathbf{G}_y\left(\mathbf{H}\mathbf{G}_x\right)^{-1}\mathbf{H}\left[\begin{array}{cc} \mathbf{F}\mathbf{W}_d & \mathbf{W}_{n^x} \end{array}\right]\left[\begin{array}{c} \mathrm{d}' \\ \mathrm{n}^{x'} \end{array}\right]}_{\mathrm{e}_1} + \underbrace{\left[\mathbf{G}_y\left(\mathbf{H}\mathbf{G}_x\right)^{-1} - \mathbf{I}\right]\mathrm{y}_s}_{\mathrm{e}_2} \tag{24}$$

In the first term of equation (24) we have extra degree of freedom, because if we pre-multiply the first term by a matrix $\mathbf{D}$, we will have

$$\mathrm{e}_1\left(\mathbf{H}\right) = \mathrm{e}_1\left(\mathbf{D}\mathbf{H}\right)$$

where $\mathbf{D}$ is any nonsingular square matrix. This follows because

$$\left(\mathbf{D}\mathbf{H}\mathbf{G}_x\right)^{-1}\mathbf{D}\mathbf{H} = \left(\mathbf{H}\mathbf{G}_x\right)^{-1}\mathbf{D}^{-1}\mathbf{D}\mathbf{H} = \left(\mathbf{H}\mathbf{G}_x\right)^{\mathbf{-1}}\mathbf{H}$$

$\mathbf{D}$ can be chosen freely without affecting $\mathrm{e}_1\left(\mathbf{H}\right)$, so we may choose it such that the last term is zero, $\mathrm{e}_2\left(\mathbf{H}\right) = 0$, corresponding to having $\mathbf{H}\mathbf{G}_x = \mathbf{G}_y$. This means that the optimal $\mathbf{H}$ can be found by minimizing the first term (e_1) in equation (24). This problem is equivalent to solving the constrained minimization problem (18) which is convex [1]. □

The optimization problem in equation (18) is expressed with open-loop gains ($\mathbf{G}_x$ and $\mathbf{G}_y$), but can also be expressed with closed-loop gains by just substituting the open-loop gains for the closed-loop gains. This can easily be shown by postmultiplying the constraint $\mathbf{H}\mathbf{G}_x = \mathbf{G}_y$ with $\mathbf{G}_y^{-1}$ on both sides of the equality, to get $\mathbf{H}\mathbf{G}_x\mathbf{G}_y^{-1} = \mathbf{H}\mathbf{G}_x^{cl} = \mathbf{I}$

2.3.1. Analytical Solution for $\mathbf{H}$

If $\tilde{\mathbf{F}} = \left[\begin{array}{cc} \mathbf{F}\mathbf{W}_d & \mathbf{W}_{\mathbf{n}^{\mathbf{x}}} \end{array}\right]$ is full rank, which is always the case if we include independent measurement noise (so that $\mathbf{W}_{n^x}$ is full rank), then we may alternatively use the analytic expression shown in the Theorem below.

One special case, when the expression for $\mathbf{H}$ applies also for $\mathbf{W}_{n^x} = 0$, is when number of disturbances are more than the number of measurements, because $\tilde{\mathbf{F}}\tilde{\mathbf{F}}^T$ remains full rank.

Theorem 5. *Under the assumption that* $\tilde{\mathbf{F}}\tilde{\mathbf{F}}^T$ *is full rank, an analytical solution for the problem (18) is*

$$\mathbf{H}_{CL}^T = \left(\tilde{\mathbf{F}}\tilde{\mathbf{F}}^T\right)^{-1} \mathbf{G}_x \left(\mathbf{G}_x^T \left(\tilde{\mathbf{F}}\tilde{\mathbf{F}}^T\right)^{-1} \mathbf{G}_x\right)^{-1} \mathbf{G}_y \tag{25}$$

Proof: The proof is in [1] and is based on first vectorizing the problem and then using standard results from constrained quadratic optimization.

This solution is equivalent to the following [21]

$$\mathbf{H}_{CL} = \mathbf{D}\left(\left(\tilde{\mathbf{F}}\tilde{\mathbf{F}}^T\right)^{-1} \mathbf{G}_x\right)^T \tag{26}$$

where

$$\mathbf{D} = \mathbf{G}_y \left(\mathbf{G}_x^T \left(\tilde{\mathbf{F}}\tilde{\mathbf{F}}^T\right)^{-1} \mathbf{G}_x\right)^{-1} \tag{27}$$

The following example shows the effect of noise for various scenarios.

2.4. Example: We consider a scalar case with one input (u)*, one disturbance* (d)*, one measurement* (x)*, one output* y

, and with the following model matrices

$$\mathbf{G}_x = \mathbf{G}_x^d = 1$$
$$\mathbf{G}_y = \mathbf{G}_y^d = 1$$
$$\mathbf{W}_u = \mathbf{W}_d = \mathbf{W}_{y_s} = 1$$

This corresponds to the case where y = x and we have $\mathbf{F} = 0$.

For case S1, Theorem 1 gives

Table 1: Optimal $\mathbf{H}$ matrix for different scenarios

$\mathbf{W}_{n^x}$	$\mathbf{H}_1$	$\mathbf{H}_2$	$\mathbf{H}_{CL}$
0	1	1	1
1	0.67	0.5	1
5	0.074	0.038	1
∞	0	0	1

$$\mathbf{Y}_1 = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}$$
$$\mathbf{X}_1 = \begin{bmatrix} 1 & 1 & \mathbf{W}_{n^x} \end{bmatrix}$$

and we find

$$\mathbf{H}_1 = \frac{2}{\mathbf{W}_{n^x}^2 + 2}$$

For case S2, Theorem 2 gives

$$\mathbf{Y}_2 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$
$$\mathbf{X}_2 = \begin{bmatrix} 1 & 0 & \mathbf{W}_{n^x} \end{bmatrix}$$

and we find

$$\mathbf{H}_2 = \frac{1}{\mathbf{W}_{n^x}^2 + 1}$$

For case S4, Equation 25 gives

$$\tilde{\mathbf{F}}\tilde{\mathbf{F}}^T = (\mathbf{W}_{n^x})^2$$

and we have $\mathbf{H}_{CL} = 1$ for all values of the measurement noise.

Table 1 shows the optimal $\mathbf{H}$ for the three cases for some values of the measurement noise. For the "monitoring" cases ($\mathbf{H}_1$ and $\mathbf{H}_2$), the optimal estimator gain $\mathbf{H}$ approaches zero when the measurement noise goes to infinity, but this does not occur for the closed-loop estimator ($\mathbf{H}_{CL}$). The reason is that the estimate $\hat{\mathrm{y}} = \mathbf{H}_{CL}\mathrm{x}_m$ is used for control, that is, u is changed such that $\hat{\mathrm{y}}$ is equal to y_s. If we used an estimator where $\mathbf{H}_{CL} \to 0$ then we would need u to go to infinity, which is not optimal.

3. Data-based Estimators

So far we have assumed that we have available a model, which are given by $\mathbf{G}_x$, $\mathbf{G}_x^d$, $\mathbf{G}_y$ and $\mathbf{G}_y^d$ in equations 8 and 9, and the expected magnitudes of disturbances and measurement noise, etc. were given by weighting matrices (e.g. $\mathbf{W}_d$ and $\mathbf{W}_n^x$). Here, we consider the data-based case where we start from the observations collected in the matrices $\mathbf{X}$ and $\mathbf{Y}$. We want to obtain a linear relationship between the data sets.

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{B}_0 \tag{28}$$

where $\mathbf{B}$ and $\mathbf{B}_0$ as optimization variables. $\mathbf{B}_0$ is normally zero if the data are centered.

3.1. Least Squares Solution

The least-square solution to this problem is

$$\mathbf{B} = \mathbf{Y}\mathbf{X}^{\dagger} \tag{29}$$

It can be seen to be a direct generalization of the "monitoring" estimators in Theorems 1,2 and 3.

3.2. Principal Component Regression (PCR) Method

A variant of Least Square is PCR. It starts with a principal component singular value analysis of the data matrix $\mathbf{X}$, to remove directions in $\mathbf{X}$ data with little information. The matrix is truncated to rank a, where a is the number of principal components, and gives $\tilde{\mathbf{X}} = \tilde{\mathbf{U}}_a\tilde{\mathbf{\Sigma}}_a\tilde{\mathbf{V}}_a^T$. The optimal estimator is then

$$\mathbf{B}_{PCR} = \mathbf{Y}\tilde{\mathbf{X}}^{\dagger}$$

where $\tilde{\mathbf{X}}^{\dagger}$ is the inverse of the truncated SVD of the matrix $\mathbf{X}$.

3.3. Partial Least Square (PLS) Method

In its general form PLS creates orthogonal score vectors (called latent vectors or components) by maximizing the covariance between different sets of variables. There are several different algorithms generating bases which all give the same predictor, when there is one $\mathbf{Y}$ variable. Rosipal and Kramer [18] present a review of the different forms. The first approach was nonlinear iterative partial least squares (NIPALS) by Wold [25]. The predictor data matrix $\mathbf{X} = [\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_r]$, containing the values of r predictors for N samples is compressed into a set of A latent variable or factor scores $\mathbf{T} = [\mathrm{t}_1, \mathrm{t}_2, \ldots, \mathrm{t}_A]$, where $a \leq r$. These factors are determined sequentially using NIPALS. The orthogonal factor scores are used to fit a set of N observations to m dependent variables $\mathbf{Y} = [\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_m]$.

There are some assumptions which are inherent in the problem definition or some in the solution procedure, which are as follows [7]:

1. Assume centered data generated according to the latent variable model

2. Weight matrix should have orthonormal column vectors

3. The number of y variables is less than the number of components ($m \leq A$)

4. Components of regressor variables and response variables are independent, i.e. diagonal expectations $\mathrm{E}\left(\mathrm{x}_k, \mathrm{x}_k^T\right) = 0$ and $\mathrm{E}\left(\mathrm{y}_k, \mathrm{y}_k^T\right) = 0$

5. The most important assumption is that the outputs and the input data have linear relationship.

The NIPALS method includes various iterative orthogonalization (deflation) processes. Di Ruscio [5] has presented a new interpretation and description of the basic PLS solution which is non-iterative (see Appendix), which is more interesting for control community. This solution can be expressed in terms of some weighting vectors only. The equivalence between this method and the NIPALS version of the PLS method is demonstrated by Elden [6] by proving that they give the same sequence of orthogonal basis vectors. The weight matrix $\mathbf{W}_a$ is of size $r \times a$ (so the number of components, a, should of course first be specified). They have first calculated the weight vectors by an orthogonalization process. The solution is parameterized as $\mathbf{B} = \mathbf{W}_a \mathrm{p}$, where the vector p is chosen to minimize the Frobenius norm of $\mathbf{Y} - \mathbf{X} \times \mathbf{B} = \mathbf{Y} - \mathbf{X} \times \mathbf{W}_a \times \mathrm{p}$ for some specified weighting matrix $\mathbf{W}_a$.

The orthogonalization process for calculating the weight vectors is not unique. It is evident that any weighting matrix defined as $\mathbf{W}_a := \mathbf{W}_a \mathbf{D}$ (where $\mathbf{D} \in \mathbb{R}^{a \times a}$is defined as a non-singular transformation matrix) can be a solution for this problem, as mentioned by Di Ruscio [5]. So, by taking the weights $\mathbf{W}_a$ from the Krylov subspace or from the space which span the Krylov subspace, the optimal weights will be found in the sense that an iterative Ordinary Least square (OLS) converges the fastest to the OLS solution, i.e. in a minimum number of iterations [5].

4. New data-based estimation

We want to use our results for the optimal model-based estimators, to derive data-based estimators. The first step is to obtain the required model to use for cases S1-S4 in Theorems 1-4.

4.1. Monitoring cases

For cases S1-S3, all the optimal estimators are on the form $\mathbf{H} = \mathbf{Y}\mathbf{X}^{\dagger}$, so we may use the data directly. The result will be identical to the conventional least squares solution, which from our derivation should be the optimal estimator for the case when there is no measurement noise for y.

4.2. Closed-loop estimator

Let us now consider the more interesting case S4, where we want to find the optimal estimator to be used for closed-loop operation. To use Theorem 4, we need to have information about $\tilde{\mathbf{F}} = \left[\begin{array}{cc} \mathbf{F}\mathbf{W}_d & \mathbf{W}_{\mathbf{n}^{\mathbf{x}}} \end{array}\right]$ and $\mathbf{G}_x \mathbf{G}_y^{-1} = \mathbf{G}_x^{cl}$.

This information can be obtained by transforming the original data in $\mathbf{Y}$ and $\mathbf{X}$, to match the "closed-loop" form as given by the matrices $\mathbf{Y}_2$ and $\mathbf{X}_2$ in (16):

$$\mathbf{Y}_2 = \begin{bmatrix} \mathbf{W}_{y_s} & 0 \end{bmatrix}$$
$$\mathbf{X}_2 = \begin{bmatrix} \mathbf{G}_x^{cl}\mathbf{W}_{y_s} & \tilde{\mathbf{F}} \end{bmatrix}$$

This may be done as follows. Collect all the experimental data in the big matrix $\mathbf{Y}_{all}$.

$$\mathbf{Y}_{all} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{X} \end{bmatrix} \tag{30}$$

Then

1. Perform a singular value decomposition on the data matrix $\mathbf{Y} = \mathbf{U\Sigma V}^T$
2. Multiply the data matrix $\mathbf{Y}_{all}\mathbf{V}$ with the unitary matrix $\mathbf{V}$ to get $\mathbf{Y}_{all}\mathbf{V}$ on the desired form

$$\mathbf{Y}_{all} = \begin{bmatrix} \mathbf{W}_{y_s} & 0 \\ \mathbf{G}_x^{cl}\mathbf{W}_{\mathbf{y_s}} & \tilde{\mathbf{F}} \end{bmatrix} \tag{31}$$

where $\tilde{\mathbf{F}} = \begin{bmatrix} \mathbf{F}\mathbf{W}_d & \mathbf{W}_{n^x} \end{bmatrix}$
Note that $\mathbf{F}$ is defined as $\left(\frac{\partial \mathrm{x}}{\partial \mathrm{d}}\right)_{\mathrm{y}=0}$
Since $\mathbf{V}$ is a unitary matrix, the magnitude of the prediction error does not change when it is multiplied by $\mathbf{V}$, so $\|\mathbf{YV} - \mathbf{HXV}\|_F = \|\mathbf{Y} - \mathbf{HX}\|_F$. This follows because the singular vectors satisfy $\mathbf{V}^T = \mathbf{V}^{-1}$, so we have

$$\mathbf{YV} = \mathbf{U\Sigma} = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_1 \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{U}_1\mathbf{\Sigma}_1 & 0 \end{bmatrix}$$

The closed-loop data-based estimator (CL) suffers from the same weakness as LS, giving poor results for ill-conditioned matrices and underdetermined systems. Performing a principal component analysis on the $\mathbf{X}$ data will remove the weaker directions containing noise resulting in a well-conditioned matrix. Then, CL can be applied to the data. We call this “truncated CL”.

5. Examples

5.1. Example 1

To investigate the performance of the estimators, they were applied to a linear approximation of a binary distillation column model - Column A [20] - subjected to different control scenarios. Full information about the model and the source codes are online [19]. There are two inputs, namely the reflux flow and the boilup, and one disturbance, which is the change in feed composition. The model for the two primary variables is

$$\mathrm{y} = \begin{bmatrix} 0.175 & -0.164 \\ 1.764 & -1.773 \end{bmatrix} \mathrm{u} + \begin{bmatrix} 0.164 \\ 1.836 \end{bmatrix} \mathrm{d} \tag{32}$$

where the primary variables are compositions of the two main components in the top and bottom products. The model for the eight measurements (temperatures) is

$$\mathrm{x}_m = \begin{bmatrix} -190.292 & 189.035 \\ -229.539 & 231.298 \\ -50.149 & 50.743 \\ -70.084 & 69.106 \\ -154.121 & 154.457 \\ -149.137 & 148.847 \\ -215.412 & 216.714 \\ -194.170 & 192.475 \end{bmatrix} \mathrm{u} + \begin{bmatrix} -203.828 \\ -244.896 \\ -51.995 \\ -71.375 \\ -170.510 \\ -164.730 \\ -232.326 \\ -205.026 \end{bmatrix} \mathrm{d} + \mathrm{n}^x \tag{33}$$

These measurements are chosen randomly from the top and bottom sections in the column. The two secondary variables, which are reflux flow and a temperature measurement from 25th tray of column, are given by

$$\mathrm{z} = \begin{bmatrix} 1 & 0 \\ -154.121 & 154.457 \end{bmatrix} \mathrm{u} + \begin{bmatrix} 0 \\ -170.510 \end{bmatrix} \mathrm{d} \tag{34}$$

The disturbance and noise variances are as below for all scenarios:

$$\mathrm{d} \sim \mathcal{N}\left(0, 0.05^2 \mathbf{I}_2\right)$$

$$\mathrm{n}^x \sim \mathcal{N}\left(0, 0.5^2 \mathbf{I}_8\right)$$

Since there is no control in the first scenario, the standard deviation in u ($\sigma \approx$ 0.05) was selected to give a small standard deviation in y. The resulting standard deviations in the primary variables for all scenarios are the same.

Table 2: Four operation scenarios

Operation	Estimator	Input variables	Variable distribution
Open-loop	$\mathbf{H}_1$	u	$\mathrm{u} \sim \mathcal{N}\left(0, 0.05^2 \mathbf{I}_2\right)$
$\mathrm{y} = \mathrm{y}_s$	$\mathbf{H}_2$	$\mathbf{G}_y^{-1}\left(\mathrm{y}_s - \mathbf{G}_y^d \mathrm{d}\right)$	$\mathrm{y}_s \sim \mathcal{N}\left(0, 0.005^2 \mathbf{I}_2\right)$
$\mathrm{z} = \mathrm{z}_s$	$\mathbf{H}_3$	$\mathbf{G}_z^{-1}\left(\mathrm{z}_s - \mathbf{G}_z^d \mathrm{d}\right)$	$z_s \sim \mathcal{N}\left(0, \left[\ 0.05^2 \quad 2^2\ \right] \mathbf{I}_2\right)$
$\hat{\mathrm{y}} = \mathrm{y}_s$	$\mathbf{H}_{CL}$	$(\mathbf{H}\mathbf{G}_x)^{-1}\left[\mathbf{H}\left(\mathbf{G}_x^d \mathrm{d} + \mathrm{n}^x\right) + \mathrm{y}_s\right]$	$\mathrm{y}_s \sim \mathcal{N}\left(0, 0.005^2 \mathbf{I}_2\right)$

For the data based estimators, calibration data was generated by drawing 32 random values for u, d, y_s and z_s with the distributions given in Table 2, and calculating the corresponding output variables x_m and y for the respective scenarios (except scenario 4). This gave one set of calibration data with 32 experiments: $\mathbf{X}$ (8×32) and $\mathbf{Y}$ (2×32). The median of the prediction error for 150 runs are used to assess the estimators' performances because noise and variation in input variables resulted in a distorted picture of estimator performance by outliers.

5.1.1. Model-based estimators

Table 3 shows the results of validation for model-based for different cases. For each case (S1, S2, S3 and S4), the matrix $\mathbf{H}$ is obtained first, using theorems 1-4. For example for case S4 we obtain

$$\mathbf{H}_{CL} = \begin{bmatrix} -0.0024 & 0.0008 \\ 0.0004 & -0.0041 \\ -0.0001 & -0.0017 \\ -0.0025 & -0.0001 \\ 0.0011 & 0.0004 \\ 0.0003 & 0.0013 \\ 0.0007 & -0.0026 \\ -0.0037 & 0.0005 \end{bmatrix}$$

Then, they were validated on the data generated randomly for each case (S1, S2, S3, S4), with the given standard deviations for n^x, u, z_s, y_s (see Table 2). The validation is done by first calculating u for the given case and then substituting into the model. The reported data in Table 3 shows the median of the prediction errors. In Table 3, the diagonal elements correspond to the optimal estimators for the intended cases. As expected, the prediction error is smallest along the diagonal. Note that the cases are not comparable along the rows because of different variances for different cases. Calibrating with one case and validating with another is mostly applicable to the last case. So, the shaded cells are actually showing the more interesting data.

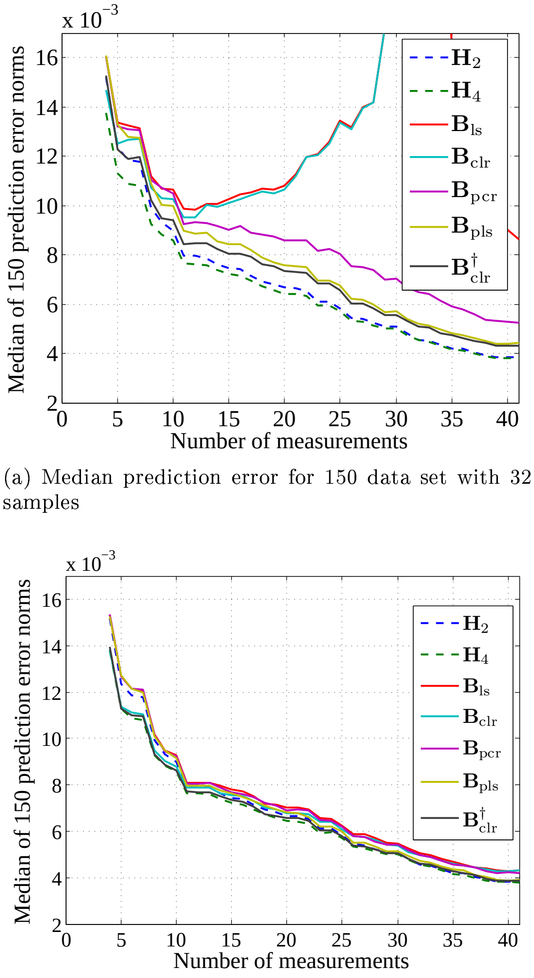
Table 3: The mean prediction error of the model-based estimators applied to four operation scenarios

Estimator	S1	S2	S3	S4
S1	0.0168	0.0248	0.0177	0.1972
S2	0.271	0.0156	0.035	0.0221
S3	0.0207	0.0224	0.0176	0.1021
S4	0.0187	0.0187	0.0187	0.0187

The prediction errors are equal for all the cases for S4 due to the constraint $\mathbf{G}_y = \mathbf{H}\mathbf{G}_x$. The closed loop estimator gives the best performance as model-based estimator. A low prediction error for case 1 when the data are calibrated on case 1 may be misleading. Case 3 is the most used arrangement in industry, where a secondary variable is controlled.

5.1.2. Data-based estimators

Figure 4 shows the estimators "closed-loop" performance with two different numbers of data. The number of regressors are increased from 8 to 41 (the total number of stages). All estimators are trained on calibration data from scenario 2 and validated on scenario 4. The performance of CL and LS was deteriorated when then the system was over-determined with low number of data. This is because they were forced to use the weak directions and assimilate noise and collinearity. Since truncated CL filters out the noise, it results in better performance. Comparing the two figures in Figure 4, we will see that if the data-based estimators are given enough data they will approach their model-based counterparts.



(a) Median prediction error for 150 data set with 32 samples

(b) Median prediction error for 150 data set with 200 samples

Figure 4: Median prediction error for two sample sizes (validated for S4)

5.2. Example 2

The next example is from a multi-component distillation column (4 components) which is simulated rigorously. The schematic of the distillation process with estimator is shown in Figure 5.

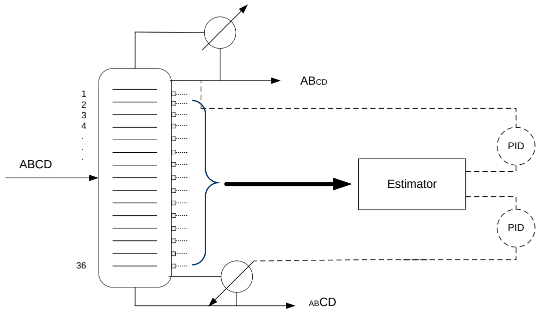


Figure 5: Schematic of the distillation process with estimator

The two lightest and the two heaviest products are supposed to be separated in the column. The feed stream is a saturated liquid mixture of methanol, ethanol, 1-propanol, 1-butanol. Disturbances are composition, flow rate, quality, Pressure in the feed stream and also condenser pressure.The composition setpoints for 1-propanol in the top ($x_{C_3\ in\ D}$) and ethanol in the bottom ($x_{C_2\ in\ B}$) of prefractionator are 0.0095 and 0.038 respectively.

Here we show how simple the closed-loop model-based estimator can be derived by choosing the right variables as manipulated variable. We can actually consider u to be any two variables from the process. For the sake of simplicity and because we can use the close-loop information of the system, we select the inputs to the estimation model to be equal to the product compositions, in our case

$$\mathrm{u} = \mathrm{y} = \begin{bmatrix} x_{C_3 in\, D} & x_{C_2 in\, B} \end{bmatrix}$$

This will make the case easier and the matrices will be as below:

$$\mathbf{G}_y = \mathbf{I}$$

$$\mathbf{G}_x^d = \mathbf{F}$$

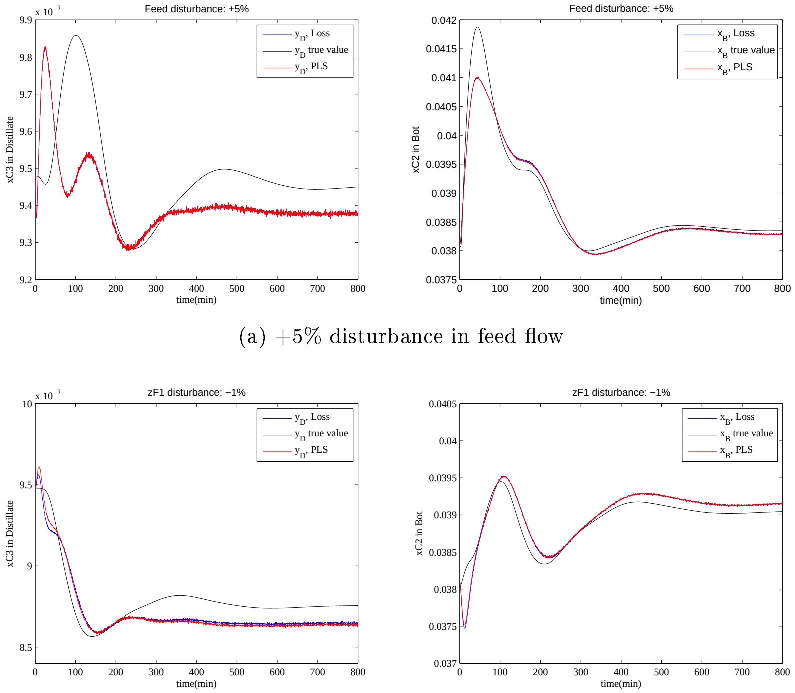
$$\mathbf{G}_y^d = 0$$

We use exactly the same information for PLS method. **X** and **Y** in PLS method are the first and second row of $\mathbf{Y}_{all}$ matrix (Equation (**??**)) respectively. We have assumed that we have temperature sensors in every 4th tray. The matrices in the following show the fitting matrices for the two methods (**B** for PLS and **H** for Loss method).

$$\mathbf{B} = \begin{bmatrix} 0.0002 & 0.0013 \\ 0.0087 & -0.0041 \\ -0.006 & 0.0068 \\ -0.0051 & 0.0003 \\ 0.0077 & -0.0096 \\ -0.0034 & 0.0124 \\ -0.0016 & 0.0049 \\ 0.0026 & -0.016 \\ -0.0031 & 0.004 \end{bmatrix}$$

$$\mathbf{H}_{CL} = \begin{bmatrix} 0.0004 & 0.0014 \\ 0.0081 & -0.0045 \\ -0.005 & 0.0074 \\ -0.0047 & 0.0006 \\ 0.0062 & -0.0104 \\ -0.003 & 0.0126 \\ -0.0013 & 0.0051 \\ 0.0024 & -0.0162 \\ -0.0028 & 0.0042 \end{bmatrix}$$

Figure 6 shows the dynamic behaviour of the model as disturbances happen and also of the estimators. It is shown that the estimated values can track the real composition very well. It should be noted that the steady state value is more in focus since the methods under study are static estimators. The dynamic behaviour can be corrected by feedback.



(a) +5% disturbance in feed flow

(b) -1% disturbance in Feed composition $z_{1,F}$

Figure 6: Estimated and model Composition values for the case with two temperature controls and with the consideration of 8 measurements

6. Discussion

6.1. Relationship to self-optimizing control

This work originated from considering the "indirect control problem" ([9]) using the "exact local method" in self-optimizing control. In "indirect control" the objective is to find a set of controlled variables $\mathrm{c} = \mathbf{H}\mathrm{x}$ such that by keeping c constant, we indirectly keep the primary variables y constant (or more specifically, at their desired setpoints y_s), in spite of disturbances d and measurement noise n^x. This can be viewed as a special case of "self-optimizing control" with cost function $\mathrm{J} = \|\mathrm{y} - \hat{\mathrm{y}}\|_2$ ([9]). We can then apply the theory that has been developed for "self-optimizing" control, which includes the "exact local method". This directly leads to the result in Theorem 4, when the "extra degrees of freedom" in $\mathbf{H}$ are selected such that $\mathrm{c} = \hat{\mathrm{y}}$. This requires some explanation. In indirect control, we adjust the inputs u by feedback to keep $\mathrm{c} = \mathbf{H}\mathrm{x} = 0$ (constant). Note that we will generate the same inputs u (for a given d and n^x), also if we keep $\mathrm{c}' = \mathbf{D}\mathrm{c} = 0$ where $\mathbf{D}$ is any invertible matrix. The matrix $\mathbf{D}$ is the so called "extra degrees of freedom" in $\mathbf{H}$.

It is clear that one good variable c = $\mathbf{H}$x to use for indirect control of y must be the estimate ŷ. However, if we look the other way around, then the optimal c will not necessarily correspond to an estimate of y (ŷ). However, there are extra degrees of freedom in selecting c = $\mathbf{H}$x, we can use these extra degrees of freedom (i.e., the $\mathbf{D}$-matrix), to make c = $\mathbf{H}$x equal to ŷ, which is in fact done when we select $\mathbf{H}$ such that $\mathbf{HG}_x = \mathbf{G}_y$ (see Theorem 4).

6.2. Comparison with work of Pannocchia and Brambilla

Our paper provides an extension of the results of Pannocchia and Brambilla ([17]) on "steady-state closed-loop consistency" to include also measurement noise. In addition, we have shown, in agreement with the results in [9], that we can always achieve "perfect consistency" for setpoint changes, that is, the use of the "extra degrees of freedom" in $\mathbf{H}$, makes it possible to always have the norm from y_s to the prediction error $(y-\hat{y})$ equal to zero, without sacrificing the norm from disturbances (d) to the prediction error. In the notation of Pannocchia and Brambilla ([17]) this means that we can always make $\epsilon_r = 0$ without sacrificing the norm of ϵ_d.

The inclusion of measurement noise is important, because this is often a critical factor. As an example, consider the estimation of the two product compositions in a distillation column ($y = \left[\begin{array}{cc} x_D & x_B \end{array}\right]$) based on temperature measurements (x = T). For a binary distillation column with constant pressure, temperature and compositions are uniquely related. So, if there were no measurement noise ($n^x = 0$), one could in theory have a perfect estimate of y by measuring the temperature at the two columns ends ($x = \left[\begin{array}{cc} T_D & T_B \end{array}\right]$), irrespective of any disturbances in feed composition or feed rate (which may affect stage efficiency). However, in practice, the estimate will be poor because of measurement error, especially for high-purity columns. For example, assume the bottom product of a methanol/water distillation column should be about 99.99% water. At 1 atm, the boiling point of this mixture will be approximately (0.9999×100 °C + 0.0001× 65 °C = 99.9965 °C, whereas the boiling point of 100% water is 100.00 °C. Thus, if we have a measurement error of more than 0.0065 °C (which we certainly will have), then the temperature measurement will be useless to infer composition. Thus, due to measurement error (n^x), we need to locate the temperature sensor away from the column end, and the optimal location can be found using the methods presented in this paper which includes measurement noise.

6.3. Measurement selection

The results presented in this paper also provide the basis for optimal measurement selection, which extends the Algorithm 1 in Pannocchia and Brambilla ([17]) to include measurement noise. For example, assume there are 10 candidate measurements, and there are 2 outputs that we want to estimate (i.e, we have 2 y's and 2 u's). Assume that we want to use 4 out of these 10 measurements. There are then 210 candidate measurement sets, and we find the best set by computing for each set the prediction error using Theorem 4. To avoid checking all sets, we can also use the branch and bound method developed by Kariwala and Cao [11].

6.4. Comparison to standard data-based estimators

In least squares regression (LS), one gets $\mathbf{B} = \mathbf{Y}\mathbf{X}^{-1}$, or more generally $\mathbf{B} = \mathbf{Y}\mathbf{X}^{\dagger}$, where $\mathbf{X}^{\dagger}$ denotes the pseudo inverse of $\mathbf{X}$. In principal component regression (PCR), one gets $\mathbf{B} = \mathbf{Y}\mathbf{X}_a^{\dagger}$ where $\mathbf{X}_a^{\dagger} = \sum_{i=1}^{a} \frac{1}{\sigma_i}\nu_i u_i^H$ denotes the pseudo inverse of $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H$ with only A principal components included. Thus, in both LS and PCR one inverts the $\mathbf{X}$-matrix, while with the new loss regression method, see equation (26), one considers only a part $\mathbf{X}_{opt}$ of the transformed $\mathbf{X}$-matrix. The proposed method seems a bit similar to PLS in that we use the data for y to affect the $\mathbf{X}$-data (we get $\mathbf{X}_{opt}$ from $\mathbf{X}$ by using the SVD of $\mathbf{Y}$).

Comparing the regression equations of the loss method and PLS, we realize that the PLS method has one more degree of freedom ($\mathbf{B}_0$). By assuming deviation variables

$$\mathbf{Y} - \mathbf{Y}_0 = \mathbf{H}\left(\mathbf{X} - \mathbf{X}_0\right) \tag{35}$$

We can write the above equation as

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{H}_0 \tag{36}$$

where $\mathbf{H}_0 = \mathbf{Y}_0 - \mathbf{H}\mathbf{X}_0$
By writing

$$\mathbf{H}_0 = diag\left(\mathbf{H}_0\right) \times 1 - vector \tag{37}$$

then the problem can be written as

$$\mathbf{Y}' = \mathbf{H}'\mathbf{X}' \tag{38}$$

where $\mathbf{H}' = \begin{bmatrix} \mathbf{H} & diag\left(\mathbf{H}_0\right) \end{bmatrix}$ and $\mathbf{X}' = \begin{bmatrix} \mathbf{X} & 1 - vector \end{bmatrix}$.

So by just adding 1's to the end of the $\mathbf{X}$-data, one can optimize to find $\mathbf{H}'$ using normal least squares, and then find $\mathbf{H}$ and $\mathbf{H}_0$. It should also be mentioned that the extra degree of freedom in PLS is mostly zero because of centering.

If we look at the general equation for $\mathbf{B}$ in PLS (equation (39)) and compare it with the $\mathbf{H}$ solution (equation (26)), we see that $\mathbf{X}_{opt}$ is a variation of $\mathbf{W}_a\mathbf{X}$. $\mathbf{X}_{opt}$ is actually $\mathbf{X}\mathbf{V}$ in Loss method, where $\mathbf{V}$ is the right singular vector. It acts as some sort of $\mathbf{W}_a$.

$$\mathbf{B}_{PLS} = \mathbf{W}_a\left(\mathbf{W}_a^T\mathbf{X}^T\mathbf{X}\mathbf{W}_a\right)^{-1}\mathbf{W}_a^T w_1 \tag{39}$$

We must assume that $\mathbf{X}_{opt}$ is full rank (invertible) to use the analytic expression in equation (26). If $\mathbf{X}_{opt}$ does not have full rank one may use some pseudo-inverse of $\mathbf{X}_{opt}$ (similar to PCR). This adds degrees of freedom to the method, which in PLS is the size of the matrix $\mathbf{W}_a$ and is specified in the first step (the number of components in PLS). However, the problem of invertibility is solved by manipulating $\mathbf{W}_a$ matrix. This is one reason which makes the PLS method less robust compared to Loss Method.

The PLS method for univariate data is optimal in the prediction error sense [5]. However, the PLS algorithm for multivariate data is not optimal in the same way as PLS1. There are reports that from the literature that the PLS solution using different approaches are not equivalent. For example de Jong's SIMPLS [4] is not equivalent to Herman Wold's NIPALS.

As mentioned before, the reports from different studies showed that PLS always give a higher coefficient of determination than PCR (Table 1 in [23]). However, some authors [3, 14, 8] have taken a closer look on the shrinkage properties of PLS and have shown that PLS nearly always can be improved in principle, so the regression method as such is not optimal in any reasonable way.

7. Conclusion

In this paper, we have introduced a new static estimator. The method is based on the loss method where it is assumed that a model is available. Four cases have been used to get the calibration data and were validated for different cases. We have considered two estimators for the case with closed-loop control of the outputs y. In case S2, we assume we have perfect measurements of y, and we use the estimator $\hat{\mathrm{y}} = \mathbf{H}_2\mathrm{x}$ for monitoring purposes. Case S4 is more practically relevant, because we assume that we have no online measurement for y, and instead we control the estimated $\hat{\mathrm{y}} = \mathbf{H}_{CL}\mathrm{x}$. Nevertheless, the cases are quite similar, and we find that the estimates $\hat{\mathrm{y}}$ are similar for cases where the estimation error is small. However, if the estimate is poor, then the differences may be significant. The simulation results showed that the best performance can be obtained by model-based estimator for "closed-loop" scenario. Also, it is shown that the data-based estimators perform at best like their model-based counterparts given that the system is linear and exact model.

Acknowledgement

David Di Ruscio and Inge Svein Helland are acknowledged for discussions about PLS.

8. Appendix

Theorem 6. *(PLS1: a non-iterative solution). Di Ruscio [5]: Given data matrix* $\mathbf{X} \in \mathbb{R}^{N\times r}$ *and* $\mathbf{Y} \in \mathbb{R}^{N}$*, the PLS solution is given by*

$$\mathbf{B}_{PLS} = \mathbf{K}_a \mathrm{p}^* r \tag{40}$$

where $\mathbf{K}_a \in \mathbb{R}^{r\times a}$ *is the reduced controllability (Krylov) matrix for the pair* $\left(\mathbf{X}^T\mathbf{X}, \mathbf{X}^T\mathbf{Y}\right)$ *defined as*

$$\mathbf{K}_a = \left[\begin{array}{ccccc} \mathbf{X}^T\mathbf{Y} & \mathbf{X}^T\mathbf{X}\mathbf{X}^T\mathbf{Y} & \left(\mathbf{X}^T\mathbf{X}\right)^2\mathbf{X}^T\mathbf{Y} & \ldots & \left(\mathbf{X}^T\mathbf{X}\right)^{a-1}\mathbf{X}^T\mathbf{Y} \end{array} \right]$$

where $1 \leq a \leq r$, *and the polynomial coefficient vector* $\mathrm{p}^* \in \mathbb{R}^a$ *is determined as the LS solution to*

$$\mathrm{p}^* = \arg\min \left\|\mathrm{V}\left(\mathrm{p}\right)\right\|_F^2 \tag{41}$$

where

$$\mathrm{V}\left(\mathrm{p}\right) = \left\|\mathbf{Y} - \mathbf{X}\mathbf{K}_a\mathrm{p}\right\|_p^2 \tag{42}$$

Hence

$$\mathrm{p}^* = \left(\mathbf{K}_a^T\mathbf{X}^T\mathbf{X}\mathbf{K}_a\right)^{-1}\mathbf{K}_a^T\mathbf{X}^T\mathbf{Y} \tag{43}$$

which gives the PLS solution

$$\mathbf{B}_{PLS} = \mathbf{K}_a\left(\mathbf{K}_a^T\mathbf{X}^T\mathbf{X}\mathbf{K}_a\right)^{-1}\mathbf{K}_a^T\mathbf{X}^T\mathbf{Y} \tag{44}$$

with the assumption that $\left(\mathbf{K}_a^T\mathbf{X}^T\mathbf{X}\mathbf{K}_a\right)^{-1}$*is non-singular for some* $1 \leq a \leq r$. *The PLS prediction of* $\mathbf{Y}$ *is given by*

$$\mathbf{Y}_{PLS} = \mathbf{X}\mathbf{K}_a\mathrm{p}^* \tag{45}$$

where p^* *is given by (43). Furthermore, the minimum is*

$$\mathrm{V}\left(\mathrm{p}^*\right) = trace\left(\mathbf{Y}^T\mathbf{Y}\right) - trace\left(\mathbf{Y}^T\mathbf{X}\mathbf{K}_a\left(\mathbf{K}_a^T\mathbf{X}^T\mathbf{X}\mathbf{K}_a\right)^{-1}\mathbf{K}_a^T\mathbf{X}^T\mathbf{Y}\right) \tag{46}$$

Proof. From Cayley-Hamilton Theorem we have that $\mathbf{X}^T\mathbf{X}$ satisfies its own characteristic equation, i.e.

$$\left(\mathbf{X}^T\mathbf{X}\right)^r + \mathrm{p}_2\left(\mathbf{X}^T\mathbf{X}\right)^{r-1} + \ldots + \mathrm{p}_r\mathbf{X}^T\mathbf{X} + \mathrm{p}_{r+1}\mathbf{I}_r = 0 \tag{47}$$

where $p_2, \ldots, p_{r+1}$are the coefficients of the characteristic polynomial $\det(\lambda\mathbf{I}_r - \mathbf{X}^T\mathbf{X})$. This can be used to form the matrix inverse

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = -\frac{1}{\mathrm{p}_{r+1}}\left(\mathrm{p}_r\mathbf{I}_r + \mathrm{p}_{r-1}\mathbf{X}^T\mathbf{X} + \ldots + \mathrm{p}_2\left(\mathbf{X}^T\mathbf{X}\right)^{r-2} + \left(\mathbf{X}^T\mathbf{X}\right)^{r-1}\right) \tag{48}$$

which is derived by post-multiplying (47) with $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$and then solving for the inverse. When this equation is substituted into the OLS solution $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ gives the truncated solution

$$\mathbf{B}\left(\mathrm{p}\right) = \mathbf{K}_a\mathrm{p} \tag{49}$$

where $\mathbf{K}_a$ is the controllability matrix and $\mathrm{p} \in \mathbb{R}^a$ is the coefficient vector. We know that $\|\mathbf{A}\|_F^2 = trace\left(\mathbf{A}^T\mathbf{A}\right) = \sum_{i=1}^m\sum_{j=1}^n a_{ij}^2$, So

$$\mathrm{V}\left(\mathrm{p}\right) = trace\left(\mathbf{Y}^T\mathbf{Y}\right) - 2trace\left(\mathrm{p}^T\mathbf{K}_a\mathbf{X}^T\mathbf{Y}\right) + trace\left(\mathrm{p}^T\mathbf{K}_a^T\mathbf{X}^T\mathbf{X}\mathbf{K}_a\mathrm{p}\right) \tag{50}$$

Letting the gradient

$$\frac{d\mathrm{V}\left(\mathrm{p}\right)}{d\mathrm{p}} = -2\mathbf{K}_a^T\mathbf{X}^T\mathbf{Y} + 2\mathbf{K}_a^T\mathbf{X}^T\mathbf{X}\mathbf{K}_a\mathrm{p} \tag{51}$$

equal to zero gives the optimal solution (43), which when substituted into (40) gives (44). Furthermore, the minimum value (46) can be found by substituting the optimal truncated polynomial coefficient into (50). □

By "Non-iterative" they meant that there is no need for any deflation (rank one reduction) process in computing the PLS solution. The following theorem is an extension of PLS1 to incorporate multivariate Y data.

Theorem 7. *(CPLS: Controllability PLS solution). Di Ruscio [5] . Given data matrix* $\mathbf{X} \in \mathbb{R}^{N\times r}$ *and* $\mathbf{Y} \in \mathbb{R}^{N\times m}$ *, the PLS solution is given by*

$$\begin{aligned} \mathbf{B}_{CPLS} = \left[\begin{array}{cccc} \mathbf{X}^T\mathbf{Y} & \left(\mathbf{X}^T\mathbf{X}\right)\mathbf{X}^T\mathbf{Y} & \dots & \left(\mathbf{X}^T\mathbf{X}\right)^{a-1}\mathbf{X}^T\mathbf{Y} \end{array}\right] \times \left[\begin{array}{c} \mathrm{p}_1\mathbf{I}_m \\ \mathrm{p}_2\mathbf{I}_m \\ \vdots \\ \mathrm{p_a}\mathbf{I}_m \end{array}\right] \\ = \left(\mathrm{p}_1\mathbf{I}_r + \mathrm{p}_2\mathbf{X}^T\mathbf{X} + \mathrm{p}_3\left(\mathbf{X}^T\mathbf{X}\right)^2 + \ldots + \mathrm{p}_a\left(\mathbf{X}^T\mathbf{X}\right)^{a-1}\right)\mathbf{X}^T\mathbf{Y} \\ = \textstyle\sum_{i=1}^a \mathrm{p}_i\left(\mathbf{X}^T\mathbf{X}\right)^{i-1}\mathbf{X}^T\mathbf{Y} \end{aligned} \tag{52}$$

where the vector of polynomial coefficients

$$\mathrm{p}^* = \left[\begin{array}{cccc} \mathrm{p}_1 & \mathrm{p}_2 & \dots & \mathrm{p}_a \end{array}\right]^T \in \mathbb{R}^a \tag{53}$$

is found from the solution to the LS problem

$$\mathrm{p}^* = \arg\min_{\mathrm{p}} \|vec\left(\mathbf{Y}\right) - \mathbf{X}_{\mathrm{p}}\mathrm{p}\|_F^2 \tag{54}$$

where $vec\left(\mathbf{Y}\right)$ *is the vectorization of the matrix* $\mathbf{Y}$*. The minimizing solution is given by*

$$\mathrm{p}^* = \left(\mathbf{X}_p^T\mathbf{X}_p\right)^{-1}\mathbf{X}_p vec\left(\mathbf{Y}\right) \tag{55}$$

where

$$\mathbf{X}_p = \left[\begin{array}{cccc} vec\left(\mathbf{X}\mathbf{X}^T\mathbf{Y}\right) & vec\left(\mathbf{X}\mathbf{X}^T\mathbf{X}\mathbf{X}^T\mathbf{Y}\right) & \dots & vec\left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{a-1}\mathbf{X}^T\mathbf{Y}\right) \end{array}\right] \\ \in \mathbb{R}^{Nm\times a} \tag{56}$$

Proof. Using that $\mathrm{V}\left(\mathrm{p}\right) = \|\epsilon\|_F^2$, where ϵ is the prediction error, gives the optimal LS solution (55) by letting the gradient $\frac{d\mathrm{V}}{d\mathrm{p}} = 0$. □

9. References

[1] Alstad, V., S. Skogestad and E. S. Hori (2009). 'Optimal measurement combinations as controlled variables'. *J.Proc. Control* **19**(1), 138–148.

[2] Bhartiya, S. and R. Whiteley, J (2001). 'Development of inferential measurements using neural networks'. *ISA Transactions* **40**, 307–323.

[3] Butler, N. A. and M. C. Denham (2000). 'The peculiar shrinkage properties of partial least squares regression'. *J. R. Stat. Soc.* **B 62**, 585–593.

[4] de Jong, S. (1993). 'Simpls: An alternative approach to partial least squares regression'. *Chemom. Intell. Labor. Systems* **18**(3), 251–263.

[5] Di Ruscio, D. (2000). 'A weighted view on the partial least-squares algorithm'. *Automatica* **36**, 831–850.

[6] Elden, L. (2004). 'Partial least-squares vs. lanczos bidiagonalization - i: Analysis of a projection method for multiple regression'. *Comput. Stat. Data Anal.* **46**, 11–31.

[7] Ergon, R. (2002). 'Noise handling capabilities of multivariate calibration methods'. *Modeling, identification and control* **23**, 259–273.

[8] Helland, I. S. (2001). 'Some theoretical aspects of partial squares regression'. *Chemom. Intell. Labor. Systems* **58**, 97–107.

[9] Hori, E. S., S. Skogestad and Alstad V. (2005). 'Perfect steady-state indirect control'. *Ind. Eng. Chem. Res.* **44**, 863–867.

[10] Joseph, B. and C. B. Brosilow (1978). 'Inferential control of processes: Part i. steady state analysis and design'. *AIChE J.* **24**(3), 485–492.

[11] Kariwala, V. and Y. Cao (2009). 'Bidirectional branch and bound for controlled variable selection. part ii: Exact local method for self-optimizing control'. *Computers Chem. Eng* **33**(8), 1402–1412.

[12] Kresta, J. V., T. E. Marlin and J. F. MacGregor (1994). 'Development of inferential process models using pls'. *Computers Chem. Eng* **18**, 597–611.

[13] Li, Q. and Cheng Shao (2008). Estimation of distillation compositions using sensitivity matrix analysis and kernel ridge regression. In 'Proceedings of the 17th IFAC World Congress'. Seoul, Korea.

[14] Lingjærde, O. C. and N. Christophersen (2000). 'Shrinkage structure of partial least squares'. *Scand. J. Stat.* **27**, 459–473.

[15] Massy, W. F. (1965). 'Principal component regression in exploratory statistical research'. *J. Amer. Statist. Assoc.* **60**, 234–246.

[16] Mejdell, T. and S. Skogestad (1991a). ‘Estimation of distillation compositions from multiple temperature measurements using partial-least-squares regression’. *Ind. Eng. Chem. Res.* **30**, 2543–2555.

[17] Pannocchia, G. and A. Brambilla (2007). ‘How auxiliary variables and plant data collection affect closed-loop performance of inferential control’. *J. Process Control* **17**, 653–663.

[18] Rosipal, R. and Nicole Kramer (2006). Overview and recent advances in partial least squares. Technical report.

[19] Skogestad, S. (1991). ‘Distillation research and models @ONLINE’.

[20] Skogestad, S. (1997). ‘Dynamics and control of distillation columns: A critical survey’. *J. Modeling, Identification and Control* **18**(3), 177–217.

[21] Skogestad, S., R. Yelchuru and J. Jäschke (2011). Optimal use of measurements for control, optimization and estimation using loss method: Summary of existing results and some new. In ‘Selected Topics on Constrained and Nonlinear Control Workbook’. STU Bratislava and NTNU Trondheim.

[22] Tronci, S., F. Bezzo, M. Barolo and R. Baratti (2005). ‘Geometric observer for a distillation column: development and experimental testing’. *Ind. Eng. Chem. Res.* **44**, 9884–9893.

[23] Wentzell, P. D. and V. Montoto (2003). ‘Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures’. *Chemom. Intell. Labor. Systems* **65**, 257–279.

[24] Wold, H. (1975*a*). *Path Models with Latent Variables: The NIPALS Approach*. Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building. Academic Press.

[25] Wold, H. (1975*b*). *Soft Modelling by Latent Variables; the Nonlinear Iterative Partial Least Squares Approach*. Academic Press, London.

[26] Yan, W, W., H. Shao, H and F. Wang, X (2004). ‘Soft sensing modeling based on support vector machine and bayesian model selection’. *Comput. Chem. Eng.* **28**, 1489–1498.

[27] Zhang, J. (2001). Inferential feedback control of distillation composition based on pcr and pls models. In ‘Proceedings of the American Control Conference’. pp. 1196–1201.