

A New Class of Model-Based Static Estimators

Maryam Ghadrhan, Chriss Grimholt, and Sigurd Skogestad*

Department of Chemical Engineering, Norwegian University of Science and Technology, N-7491 Trondheim, Norway

ABSTRACT: Static estimators are commonly used as “soft sensors” in the process industry. The performance of the estimators is dependent on whether it is used for monitoring (open-loop) or for closed-loop control applications. In this work, we propose to design estimators that 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 formulas for computing the optimal static estimator. We also compare the results with standard existing estimators (e.g., partial least-squares (PLS)). The approach is applied to estimation of product composition in a distillation column from a combination of temperature measurements.

1. INTRODUCTION

In a chemical plant, there are usually a large number of sensors that are used to monitor and control processes. However, some process variables (e.g., composition) may be too difficult or expensive to measure online. Estimators, which are 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 (multivariate regression,^{1,2} artificial neural networks,³ support vector machine regression,⁴ etc.).

Principle component regression (PCR)⁵ and partial least-squares (PLS)⁶ 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 Wenzell and Montoto in 2003 compared these two methods, covering both experimental and simulation studies.⁷ In short, the advantage of PLS is that the method obtains a small prediction error with fewer principal components than for PCR. 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 co-workers.⁸ Let \tilde{u} represent the vector of independent variables, including the inputs u and the disturbances d ($\tilde{u} = [u \ d]$). Let x represent the process measurements and y the variables that we want to estimate. Let the linear static model in deviation variables be

$$x = X\tilde{u} \quad (1)$$

$$y = Y\tilde{u} \quad (2)$$

The “Brosilow” estimator is then simply the following least-squares estimate of y :

$$\hat{y} = Hx_m \quad (3)$$

where x_m are the measured variables,

$$H = YX^\dagger \quad (4)$$

and X^\dagger is the pseudo-inverse of the matrix X .

Joseph and Brosilow⁸ have discussed some of the weaknesses of this estimator. For “ill-conditioned” plants with a large condition number of X , they find that the estimate may be improved in some cases by removing measurements, because this reduces the condition number. However, removing measurements cannot be the optimal way of dealing with such problems, because we are discarding information. This is also clear when we consider the popular “data-based” regression estimators, such as PLS regression,⁹ 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 means that we handle, in an optimal manner, the “high condition number problem”, which has been a major concern in previous work.^{1,8,10–12} The derivation is straightforward, but, surprisingly, it seems it has not been presented previously.

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, this is a shortcoming of most existing data-based estimators. In the paper, we derive the 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 measured variables y at a given set point y_s . It is similar to case S1, except that the set point y_s takes the role of the inputs. Case S3 is a generalization where we control the measured variable z . Cases S1, S2, and S3 are practically relevant if the estimator is used only for monitoring, because the estimate \hat{y} is not used for control. Finally, case S4 is the relevant case when we use the estimator in closed-loop mode (for control purposes). Whereas the optimal estimators for cases for cases S1, S2, and S3 are least-squares estimators with a similar

Special Issue: John MacGregor Festschrift

Received: February 19, 2013

Revised: June 25, 2013

Accepted: July 1, 2013

Published: July 1, 2013

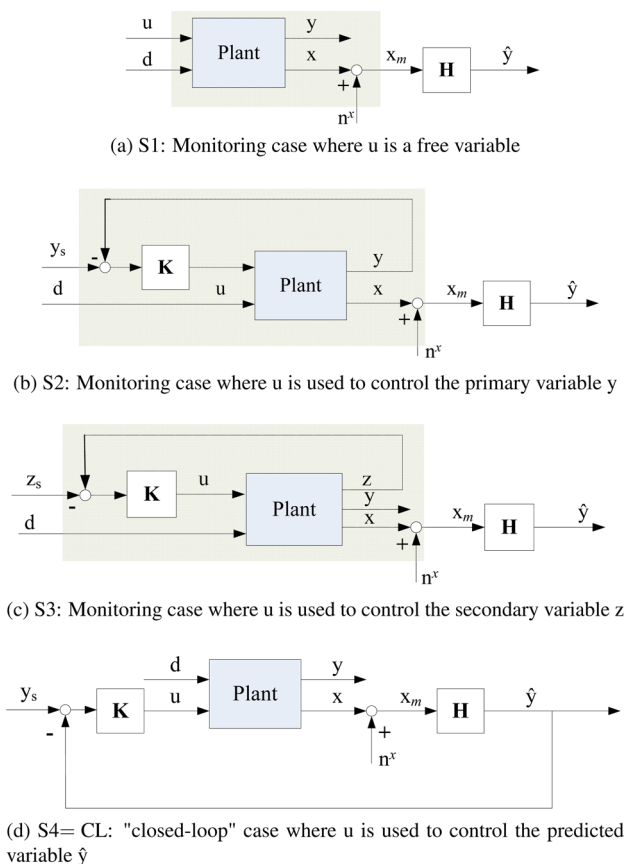


Figure 1. Block diagrams for different cases.

structure to the Brosilow estimator in eq 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 control¹³ 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 based on optimal models to derive new data-based estimators. Finally, in sections 5 and 6, we compare the new static estimators with previous work, including the Brosilow estimator and regression-based estimators on distillation case studies.

2. DERIVATION OF MODEL-BASED STATIC ESTIMATORS

2.1. Problem Definition. We define the following variables:

- u : inputs (degrees of freedom); these may include set points to 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; $\dim(z) = \dim(u)$

All variables are assumed to be deviation variables (away from the nominal or centered values). In this section, we derive optimal "open-loop" and "closed-loop" static estimators. By "optimal", it is meant that, for a linear estimator of the form

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

we want to minimize the expected prediction error:

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

The measurement signals (x_m), which are corrupted by measurement noise n^x , are given as

$$x_m = x + n^x \tag{7}$$

We use linear static models for the primary variables y , measurements x , and secondary variables z (see Figure 2):

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

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

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

In terms of the notation used for the Brosilow inferential estimator in eq 1, we have

$$\mathbf{X} = [\mathbf{G}_x \ \mathbf{G}_x^d] \tag{11}$$

$$\mathbf{Y} = [\mathbf{G}_y \ \mathbf{G}_y^d] \tag{12}$$



Figure 2. Block diagram for the linearized plant.

In addition, the expected magnitude of the independent variables for each case (see Figure 1) is quantified by weighting matrices ($\mathbf{W}_w, \mathbf{W}_d, \mathbf{W}_n, \mathbf{W}_y, \mathbf{W}_z$), 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 $\hat{y} = \mathbf{H}x_m$ are used for monitoring (that is, they are not used for control purposes). Note 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, which 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 $y = y_s$

Case S3: Predicting primary variables from a system where the inputs u are used to control the secondary variables z , i.e., $z = z_s$

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

2.2.1. Case S1. Case S1 is the direct extension of the Brosilow estimator to include noise. To determine the optimal estimator for open-loop operation, the prediction error must 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 eqs 5–9, and considering that u is a free variable, the prediction error can be expressed as

$$e(\mathbf{H}) = [(\mathbf{G}_y - \mathbf{H}\mathbf{G}_x) \ (\mathbf{G}_y^d - \mathbf{H}\mathbf{G}_x^d) \ -\mathbf{H}] \begin{bmatrix} u \\ d \\ n^x \end{bmatrix} \tag{13}$$

Proof. An expression of \hat{y} as a function of u , d , and n^x can be obtained by substituting eqs 7 and 9 into eq 5.

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

Using the definition of prediction error and substituting the expression for \hat{y} , we will have

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

which is the same as eq 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 in the forms*

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

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

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

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

$$u' \sim \mathcal{N}(0, 1)$$

$$d' \sim \mathcal{N}(0, 1)$$

$$n^{x'} \sim \mathcal{N}(0, 1)$$

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

$$e = \underbrace{\begin{bmatrix} (\mathbf{G}_y - \mathbf{H}\mathbf{G}_x)\mathbf{W}_u & (\mathbf{G}_y^d - \mathbf{H}\mathbf{G}_x^d)\mathbf{W}_d & -\mathbf{H}\mathbf{W}_{n^x} \end{bmatrix}}_{\mathbf{M}(\mathbf{H})} \begin{bmatrix} u' \\ d' \\ n^{x'} \end{bmatrix}$$

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

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

Proof. Let

$$\tilde{d} = \begin{bmatrix} u' \\ d' \\ n^{x'} \end{bmatrix}$$

Then, $e = \mathbf{M}\tilde{d}$, and noting that $\|e\|_2 = \text{tr}(ee^T)$, the expected value of the 2-norm of the prediction error can be written as

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

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

Since

$$\begin{bmatrix} u' \\ d' \\ n^{x'} \end{bmatrix} \sim \mathcal{N}(0, \mathbf{I}_{n_u+n_d+n_x})$$

by substituting the normal distribution in the definition of the expected value, we have

$$\mathbf{E}[\tilde{d}\tilde{d}^T] = \text{Var}(\tilde{d})$$

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

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

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*

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

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

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

when u is a free variable, is

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

where \mathbf{X}_1^\dagger is the pseudo-inverse of \mathbf{X}_1 , and

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

If \mathbf{X}_1 has full column rank, we have $\mathbf{X}_1^\dagger = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T$. If \mathbf{X}_1 has full row rank, we have $\mathbf{X}_1^\dagger = (\mathbf{X}_1\mathbf{X}_1^T)^{-1}\mathbf{X}_1$. 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 $\|e(\mathbf{H})\|_2$ is equivalent to minimizing $\|\mathbf{M}(\mathbf{H})\|_F^2$ 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 a least-squares problem with a 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

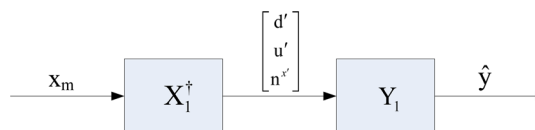


Figure 3. Interpretation of Theorem 1.

include noise. Note that the elements in Y_1 corresponding to n^x is zero.

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 set points y_s . This means that y_s replaces u as independent variables. It is assumed that $\dim(y) = \dim(u)$.

Theorem 2. The optimal “open-loop” estimator H for closed-loop operation, where the degrees of freedom u are adjusted such that the primary variables y are kept at the set points y_s , which minimizes the variance of the prediction error $y - \hat{y}$ for normally distributed set-point changes, disturbances, and noise (of magnitudes $W_{y'}, W_d$ and W_{n^s} , respectively) is

$$H_2 = Y_2 X_2^\dagger$$

where

$$Y_2 = [W_{y_s} \ 0 \ 0]$$

$$X_2 = \left[\begin{matrix} G_x^{cl} W_{y_s} & F W_d & W_{n^x} \end{matrix} \right] \tag{16}$$

where $G_x^{cl} = G_x G_y^{-1}$ and $F = G_x^d - G_x G_y^{-1} G_y^d$.

Proof. We assume that u is used to keep $y = y_s$ (with no control error). Solving eq 8, with respect to u when $y = y_s$ gives

$$u = G_y^{-1} y_s - G_y^{-1} G_y^d d$$

By combining eqs 9, 7, and 5 with the above equation, the following expression for y as an explicit function of y_s , d , and n^x is obtained:

$$\hat{y} = H [G_x G_y^{-1} y_s + (G_x^d - G_x G_y^{-1} G_y^d) d + n^x]$$

Here, $(G_x^d - G_x G_y^{-1} G_y^d) = (\partial x / \partial d)_{y=y_s}$ is the optimal sensitivity F ,¹³ and $G_x G_y^{-1} = (\partial x / \partial y_s)_d$ is known as the closed-loop gain G_x^{cl} . Therefore, the above equation becomes

$$\hat{y} = H [G_x^{cl} y_s + F d + n^x]$$

With the assumption that $y = y_s$, the prediction error becomes

$$e = y - \hat{y} = \left[\begin{matrix} (I - H G_x^{cl}) & -H F & -H \end{matrix} \right] \begin{bmatrix} y_s \\ d \\ n^x \end{bmatrix}$$

Proceeding analogous to Lemmas 1 and 2 and Theorem 1, using $y_s = W_{y'} y'_s$ results in the given preposition. ■

2.2.3. *Case S3.* The following theorem generalizes Theorems 1 and 2.

Theorem 3. The optimal “open-loop” estimator H for closed-loop operation where the degrees of freedom u are adjusted such that the secondary variables z are kept at the set points z_s , that minimizes the variance of the prediction error $y - \hat{y}$ for normally distributed set-point changes, disturbances, and noise (of magnitudes $W_{z'}, W_\delta$ and W_{n^s} , respectively) is

$$H_3 = Y_3 X_3^\dagger$$

where

$$Y_3 = [G_y^{cl} W_{z_s} \ F' W_\delta \ 0]$$

$$X_3 = [G_x^{cl} W_{z_s} \ F' W_\delta \ W_{n^x}]$$

where $G_y^{cl} = G_y G_z^{-1}$, $G_x^{cl} = G_x G_z^{-1}$, $F G_y' = G_y^d - G_y G_z^{-1} G_z^d$ and $F_x' = G_x^d - G_x G_z^{-1} G_z^d$.

Proof. We assume that u is used to keep $z = z_s$ (with no control error). Solving eq 10 with respect to u when $z = z_s$ gives

$$u = G_z^{-1} z_s - G_z^{-1} G_z^d d$$

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

$$y = \underbrace{G_y G_z^{-1} z_s}_{G_y^{cl}} + \underbrace{(G_y^d - G_y G_z^{-1} G_z^d)}_{F_y'} d$$

Introducing the optimal sensitivity F_y' and the closed-loop gain G_y^{cl} , we get

$$y = G_y^{cl} z_s + F_y' d$$

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

$$\hat{y} = H \left[\underbrace{G_x G_z^{-1} z_s}_{G_x^{cl}} + \underbrace{(G_x^d - G_x G_z^{-1} G_z^d)}_{F_x'} d + n^x \right]$$

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

$$e(H) = \left[\begin{matrix} (G_y^{cl} - H G_x^{cl}) & (F_y' - H F_x') & -H \end{matrix} \right] \begin{bmatrix} z_s \\ d \\ n^x \end{bmatrix}$$

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: setting $z = u$ gives Theorem 1 and setting $z = y$ gives Theorem 2.

2.3. The “Closed-Loop” Estimator (Case S4). In this section, we derive an expression for the optimal estimator under the assumption that the prediction $\hat{y} = H x_m$ is used for controlling the primary variables, that is, we have $\hat{y} = y_s$ (assuming integral action in the controller). It is assumed that $\dim(y) = \dim(u)$.

2.3.1. *Optimal Solution for H.* **Theorem 4.** The optimal “closed-loop” estimator H (denoted H_{CL}), following the linear relationship

$$\hat{y} = H x_m \tag{17}$$

which minimizes the variance of the prediction error

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

for normally distributed sets of d , n^x , and y_s (of magnitudes W_δ , W_{n^s} , and $W_{y'}$, respectively) assuming that the degrees of freedom u are adjusted to keep the prediction at the set point ($\hat{y} = y_s$), is

$$H_{CL} = \arg \left(\min_H \| H [F W_\delta \ W_{n^x}] \|_F \right) \tag{18a}$$

subject to

$$H G_x = G_y \tag{18b}$$

where the sensitivity matrix F is defined as

$$F = \left(\frac{\partial x}{\partial d} \right)_{y=0} = G_x^d - G_x G_y^{-1} G_y^d \tag{18c}$$

Comment: Note that eq 18b is equivalent to $H G_x^{cl} = I$.

Proof. An expression for the prediction as an explicit function of u , d , and n^x is achieved by combining eqs 9, 7, and 5 to get

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

Using a controller with integral action, the prediction \hat{y} is held at the set points y_s by manipulating u . Solving eq 19 with respect to u when $\hat{y} = y_s$, gives

$$u = -(\mathbf{H}\mathbf{G}_x)^{-1}\mathbf{H}(\mathbf{G}_x^d d + n^x) + (\mathbf{H}\mathbf{G}_x)^{-1}y_s \tag{20}$$

and inserting this into eq 8 yields

$$y = -\mathbf{G}_y(\mathbf{H}\mathbf{G}_x)^{-1}\mathbf{H}[\mathbf{F}d + n^x] + \mathbf{G}_y(\mathbf{H}\mathbf{G}_x)^{-1}y_s \tag{21}$$

where $\mathbf{F} = (\mathbf{G}_x^d - \mathbf{G}_x\mathbf{G}_y^{-1}\mathbf{G}_y^d)$ is the optimal sensitivity. Inserting the expression for y into the prediction error e , remembering that the prediction is kept at the set point ($\hat{y} = y_s$), gives

$$\begin{aligned} e &= y - \hat{y} \\ &= y - y_s \\ &= -\mathbf{G}_y(\mathbf{H}\mathbf{G}_x)^{-1}\mathbf{H}(\mathbf{F}d + n^x) + [\mathbf{G}_y(\mathbf{H}\mathbf{G}_x)^{-1} - \mathbf{I}]y_s \end{aligned} \tag{22}$$

Introducing normalized (weighted) variables gives

$$e = \underbrace{-\mathbf{G}_y(\mathbf{H}\mathbf{G}_x)^{-1}\mathbf{H} \begin{bmatrix} \mathbf{F}\mathbf{W}_d & \mathbf{W}_{n^x} \end{bmatrix}}_{e_1(\mathbf{H})} \begin{bmatrix} d \\ n^x \end{bmatrix} + \underbrace{[\mathbf{G}_y(\mathbf{H}\mathbf{G}_x)^{-1} - \mathbf{I}]y_s}_{e_2(\mathbf{H})} \tag{23}$$

In the first term of eq 23, we have an extra degree of freedom, because if we premultiply \mathbf{H} by a matrix \mathbf{D} , we will have

$$e_1(\mathbf{H}) = e_1(\mathbf{D}\mathbf{H})$$

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

$$\begin{aligned} (\mathbf{D}\mathbf{H}\mathbf{G}_x)^{-1}\mathbf{D}\mathbf{H} &= (\mathbf{H}\mathbf{G}_x)^{-1}\mathbf{D}^{-1}\mathbf{D}\mathbf{H} \\ &= (\mathbf{H}\mathbf{G}_x)^{-1}\mathbf{H} \end{aligned}$$

Since \mathbf{D} can be chosen freely without affecting $e_1(\mathbf{H})$, we may choose it such that the last term is zero ($e_2(\mathbf{H}) = 0$), 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 eq 23, subject to the constraint $\mathbf{H}\mathbf{G}_x = \mathbf{G}_y$. This problem is equivalent to solving the constrained minimization problem (18), which is convex. ■

Comment: The optimization problem in eq 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 post-multiplying the constraint $\mathbf{H}\mathbf{G}_x = \mathbf{G}_y$ with \mathbf{G}_y^{-1} on both sides of the equality, to get $\mathbf{H}\mathbf{G}_x\mathbf{G}_y^{-1} = \mathbf{H}\mathbf{G}_x^d = \mathbf{I}$.

2.3.2. Analytical Solution for H. If $\tilde{\mathbf{F}} \triangleq [\mathbf{F}\mathbf{W}_d \ \mathbf{W}_{n^x}]$ 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 in Theorem 5.

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

$$\mathbf{H}_{CL}^T = (\tilde{\mathbf{F}}\tilde{\mathbf{F}}^T)^{-1}\mathbf{G}_x(\mathbf{G}_x^T(\tilde{\mathbf{F}}\tilde{\mathbf{F}}^T)^{-1}\mathbf{G}_x)^{-1}\mathbf{G}_y \tag{24}$$

Proof. The proof is in the paper written by Alstad et al.¹³ and is based on first vectorizing the problem and then using standard results from constrained quadratic optimization. ■

Remark 1. One special case, when the expression for \mathbf{H} in eq 24 applies also for $\mathbf{W}_{n^x} = 0$, is when there are more independent terms than measurements, because $\tilde{\mathbf{F}}\tilde{\mathbf{F}}^T$ then remains full rank.¹³

Remark 2. The solution (eq 24) is equivalent to the following:¹⁴

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

where

$$\mathbf{D} = \mathbf{G}_y(\mathbf{G}_x^T(\tilde{\mathbf{F}}\tilde{\mathbf{F}}^T)^{-1}\mathbf{G}_x)^{-1} \tag{26}$$

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

2.4. Example 1. We consider a scalar case with one input (u), one disturbance (d), one measurement (x), one output y , and with the following model matrices:

$$\begin{aligned} \mathbf{G}_x &= \mathbf{G}_x^d = 1 \\ \mathbf{G}_y &= \mathbf{G}_y^d = 1 \\ \mathbf{W}_u &= \mathbf{W}_d = \mathbf{W}_y = 1 \end{aligned}$$

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

For case S1, Theorem 1 gives

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

and we find

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

For case S2, Theorem 2 gives

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

and we find

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

For case S4, eq 24 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 \mathbf{W}_{n^x} .

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

Table 1. Optimal H Matrix for Different Values of the Measurement Noise in Example 1

\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

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, $\hat{y} = \mathbf{H}_{CL}\mathbf{X}_m$ is used for control; that is, u is changed such that \hat{y} is equal to y_s . If we used an estimator where $\mathbf{H}_{CL} \rightarrow 0$, then we would need u to go toward infinity, which is not optimal.

3. EXISTING DATA-BASED ESTIMATORS

So far, we have assumed that we have available a model, which was given by G_x , G_x^d , G_y , and G_d^y in eqs 8 and 9, and the expected magnitudes of disturbances and measurement noise, etc. were given by weighting matrices (e.g., W_d and W_n^x). Here, we consider the data-based case where we start from the observations collected in the matrices X and Y . We want to obtain a linear relationship between the datasets.

$$Y = XB + B_0 \quad (27)$$

where B and B_0 as optimization variables. B_0 is normally zero if the data are centered. The resulting estimator is $\hat{y} = B x_m$. Note that we use B for a data-based estimator to distinguish it from H for a model-based estimator.

3.1. Least-Squares Solution. The least-squares solution to this problem is

$$B = YX^\dagger \quad (28)$$

It can be seen to be a direct extension of the “monitoring” estimators in Theorems 1, 2, and 3.

3.2. Principal Component Regression (PCR) Method. A variant of least squares is PCR. It starts with a principal component singular value analysis of the data matrix X , to remove directions in X data with little information. The matrix is truncated to rank a , where a is the number of principal components, and gives $\tilde{X} = \tilde{U}_a \tilde{S}_a \tilde{V}_a^T$. The optimal estimator is then

$$B_{\text{PCR}} = Y\tilde{X}^\dagger$$

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

3.3. Partial Least-Squares (PLS) Method. In its general form, partial least-squares (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 Y variable. Rosipal et al. present a review of the different forms.¹⁵ The first approach was nonlinear iterative partial least-squares (NIPALS), reported by Wold et al.¹⁶ The predictor data matrix, $X = [x_1, x_2, \dots, x_r]$, containing the values of r predictors for N samples is compressed into a set of a latent variable or factor scores $T = [t_1, t_2, \dots, t_a]$, 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 $Y = [y_1, y_2, \dots, y_m]$.

There are some assumptions that are inherent in the problem definition or some in the solution procedure, which are as follows:¹⁷

- (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 measurement variables and response variables are independent, i.e., diagonal expectations $E(x_k x_k^T) = 0$ and $E(y_k y_k^T) = 0$
- (5) The most important assumption is that the outputs and the input data have a linear relationship

The NIPALS method includes various iterative orthogonalization (deflation) processes. Di Ruscio has presented a new interpretation and description of the basic PLS solution that is noniterative, which is more interesting for the control community.¹⁸ 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 et al. by proving that they give the same sequence of orthogonal basis vectors.¹⁹ The weight matrix 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 parametrized as $B = W_a p$, where the vector p is chosen to minimize the Frobenius norm of $Y - X \times B = Y - X \times W_a \times p$ for some specified weighting matrix W_a .

The orthogonalization process for calculating the weight vectors is not unique. It is evident that any weighting matrix defined as $W_a := W_a D$ (where $D \in \mathbb{R}^{a \times a}$ is defined as a nonsingular transformation matrix) can be a solution for this problem, as mentioned by Di Ruscio.¹⁸ So, by taking the weights 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-squares (OLS) method converges the fastest to the OLS solution, i.e., within a minimum number of iterations.¹⁸

4. NEW DATA-BASED ESTIMATOR

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 $H = YX^\dagger$, 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 determine the optimal estimator to be used for closed-loop operation. To use Theorem 4, we need to have information about $\tilde{F} = [FW_d W_n^x]$ and $G_x G_y^{-1} = G_x^d$.

This information can be obtained by transforming the original data in Y and X , to match the “closed-loop” form as given by the matrices Y_2 and X_2 in (16):

$$Y_2 = [W_y \quad 0]$$

$$X_2 = [G_x^d W_y \quad \tilde{F}]$$

This may be done as follows. Collect all the experimental data in the big matrix Y_{all} .

$$Y_{\text{all}} = \begin{bmatrix} Y \\ X \end{bmatrix} \quad (29)$$

Then,

- (1) Perform a singular value decomposition on the data matrix $Y = U\Sigma V^T$.
- (2) Multiply the data matrix Y_{all} with the unitary matrix V to get $Y_{\text{all}}V$ in the desired form:

$$Y_{\text{all}}V = \begin{bmatrix} W_y & 0 \\ G_x^d W_y & \tilde{F} \end{bmatrix} \quad (30)$$

where $\tilde{F} = [FW_d W_n^x]$. \tilde{F} is denoted by X_{opt} in the following.

Note that F is defined as $(\partial x / \partial d)_{y=0}$. Since V is a unitary matrix, the magnitude of the prediction error does not change when it is

multiplied by V , so $\|YV - HXV\|_F = \|Y - HX\|_F$. This follows because the singular vectors satisfy $V^T = V^{-1}$, so we have

$$YV = U\Sigma = [U_1 \ U_2] \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} = [U_1 \Sigma_1 \ 0]$$

G_x^{cl} can be easily calculated. To determine G_x and G_y , which are needed to calculate the optimal H matrix (denoted by B_{CL} in the following), we assume that the degrees of freedom are chosen to be the primary variables. This will result in $G_y = I$.

The resulting closed-loop data-based estimator obtained from Theorem 4 or 5 (denoted B_{CL} to distinguish it from the model-based estimator H_{CL}) suffers from the same weakness as OLS, giving poor results for ill-conditioned matrices and under-determined systems. Performing a principal component analysis on the X data will remove the weaker directions containing noise, resulting in a well-conditioned matrix. Then, closed-loop data-based estimator can be applied to the data. We call this "truncated closed-loop estimator" (B_{CL}^\dagger).

5. EXAMPLES

5.1. Example 2. 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 cases. Full information about the model and the source codes is given online. There are two inputs, namely, the reflux flow and the boilup, and one disturbance, which is the change in feed composition. The linearized model for open-loop system for the two primary variables is

$$y = \begin{bmatrix} 0.8754 & -0.8618 \\ 1.0846 & -1.0982 \end{bmatrix} u + \begin{bmatrix} 0.8812 \\ 1.1188 \end{bmatrix} d \tag{31}$$

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

$$x_m = \begin{bmatrix} -64.665 & 65.413 \\ -171.884 & 173.569 \\ -226.276 & 227.842 \\ -130.878 & 130.911 \\ -195.132 & 193.623 \\ -142.092 & 140.419 \\ -55.816 & 55.013 \\ -11.818 & 11.634 \end{bmatrix} u + \begin{bmatrix} -67.174 \\ -180.728 \\ -242.622 \\ -146.618 \\ -207.430 \\ -146.928 \\ -56.667 \\ -11.897 \end{bmatrix} d + n^x \tag{32}$$

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

$$z = \begin{bmatrix} 1 & 0 \\ -195.132 & 193.623 \end{bmatrix} u + \begin{bmatrix} 0 \\ -207.4297 \end{bmatrix} d \tag{33}$$

This means that the reflux flow rate and the temperature measurement from the 25th tray are controlled, and their set points are the degrees of freedom. The disturbance and noise variances are as below for all cases:

$$d \sim \mathcal{N}(0, 0.05^2 I_2)$$

$$n^x \sim \mathcal{N}(0, 0.5^2 I_8)$$

Since there is no control in the first case, 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 are the same for all cases.

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

Table 2. Generation of Operation Data for Example 2

operation	estimator	input variables	variable distribution
open-loop	S1	u	$u \sim \mathcal{N}(0, 0.08^2 I_2)$
$y = y_s$	S2	$G_y^{-1}(y_s - G_y^d d)$	$y_s \sim \mathcal{N}(0, 0.005^2 I_2)$
$z = z_s$	S3	$G_z^{-1}(z_s - G_z^d d)$	$z_s \sim \mathcal{N}(0, [0.05^2 \ 2^2] I_2)$
$\hat{y} = y_s$	S4	$(HG_x)^{-1}[H(G_x^d d + n^x) + y_s]$	$y_s \sim \mathcal{N}(0, 0.005^2 I_2)$

output variables x_m and y for the respective cases (except case S4). This gave one set of calibration data with 32 experiments: X (8×32) and 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,

Table 3. Mean Prediction Error of the Model-Based Estimators Applied to the Four Operation Cases

estimator	Mean Prediction Error			
	case S1	case S2	case S3	case S4
H_1	0.0168	0.0248	0.0177	0.1972
H_2	0.271	0.0156	0.035	0.0221
H_3	0.0207	0.0224	0.0176	0.1021
$H_4 = H_{CL}$	0.0187	0.0187	0.0187	0.0187

S2, S3, and S4), the matrix H is obtained first, using Theorems 1–4. For example, for case S4, we obtain

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

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, and, 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 boldface cells are actually showing the more interesting data.

The prediction errors are equal for all the operation cases for \mathbf{H}_{CL} because of the constraint $\mathbf{G}_y = \mathbf{H}\mathbf{G}_x$. The closed-loop estimator generally gives the best performance.

5.1.2. *Data-Based Estimators.* Table 4 shows the results of validation for 15 data-based estimators. As mentioned in the

Table 4. Mean Prediction Error of the Data-Based Estimators Applied to the Four Operation Scenarios

training data	estimator	Mean Prediction Error (Validation)			
		case S1	case S2	case S3	case S4
S1	$\mathbf{B}_{LS,1}$	0.017	0.019	0.018	0.02
S2	$\mathbf{B}_{LS,2}$	0.316	0.016	0.061	0.173
S3	$\mathbf{B}_{LS,3}$	0.077	0.022	0.017	0.054
S1	$\mathbf{B}_{PCR,1}$	0.017	0.018	0.017	0.018
S2	$\mathbf{B}_{PCR,2}$	0.379	0.015	0.069	0.192
S3	$\mathbf{B}_{PCR,3}$	0.091	0.023	0.016	0.065
S1	$\mathbf{B}_{PLS,1}$	0.016	0.018	0.017	0.018
S2	$\mathbf{B}_{PLS,2}$	0.352	0.014	0.067	0.192
S3	$\mathbf{B}_{PLS,3}$	0.077	0.021	0.016	0.055
S1	$\mathbf{B}_{CL,1}$	0.018	0.02	0.018	0.020
S2	$\mathbf{B}_{CL,2}$	0.132	0.018	0.028	0.067
S3	$\mathbf{B}_{CL,3}$	0.077	0.022	0.017	0.053
S1	$\mathbf{B}_{CL,1}^\dagger$	0.017	0.019	0.017	0.019
S2	$\mathbf{B}_{CL,2}^\dagger$	0.130	0.016	0.028	0.066
S3	$\mathbf{B}_{CL,3}^\dagger$	0.088	0.024	0.016	0.061

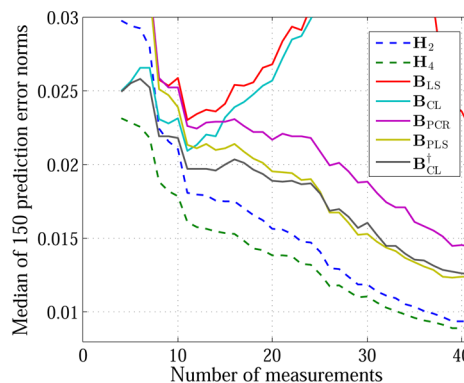
previous section, the boldface cells are the more interesting data. There is no closed-loop training data (S4) because these cannot be generated before the estimator is found.

Figure 4 shows the “closed-loop” performance with two different datasets. The number of measurements is increased from 8 to 41 (the total number of stages). All estimators are trained on calibration data from case S2 (y controlled) and validated on case S4 (\hat{y} controlled). The performance of new closed-loop estimator (\mathbf{B}_{CL} in Figure 4) and the OLS estimator (\mathbf{B}_{LS} in Figure 4) deteriorated when then the system was overdetermined with a low number of data. This is because they were forced to use the weak directions and assimilate noise and collinearity. Since the truncated closed-loop estimator (\mathbf{B}_{CL}^\dagger in Figure 4) filters out the noise, it results in better performance. Comparing the two figures in Figure 4, we see that if the data-based estimators are given enough training data they approach their model-based counterparts ($\mathbf{H}_2, \mathbf{H}_4$).

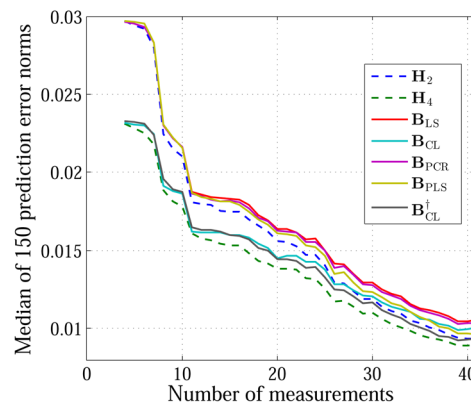
5.2. **Example 3.** The next example is from a multicomponent distillation column (four components), which is simulated rigorously. The schematic of the distillation process with estimator is shown in Figure 5.

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, and 1-butanol. Disturbances are composition, flow rate, quality, pressure in the feed stream, and condenser pressure. The composition set points for 1-propanol in the top ($x_{C_3 \text{ in D}}$) and ethanol in the bottom ($x_{C_2 \text{ in B}}$) of the 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 input



(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 training sample sizes ((a) 32 training samples and (b) 200 training samples) (validated for case S4).

set u . 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,

$$u = y = [x_{C_3 \text{ in D}} \quad x_{C_2 \text{ in B}}]$$

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

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

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

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

We use exactly the same information for the PLS method. \mathbf{X} and \mathbf{Y} in the PLS method are, respectively, the first and second row of the \mathbf{Y}_{all} matrix (eq 29). We have assumed that we have temperature sensors in every fourth tray. The matrices in the following show the fitting matrices for the two methods (\mathbf{H}_{PLS} for PLS and \mathbf{H}_{CL} is from eq 24).

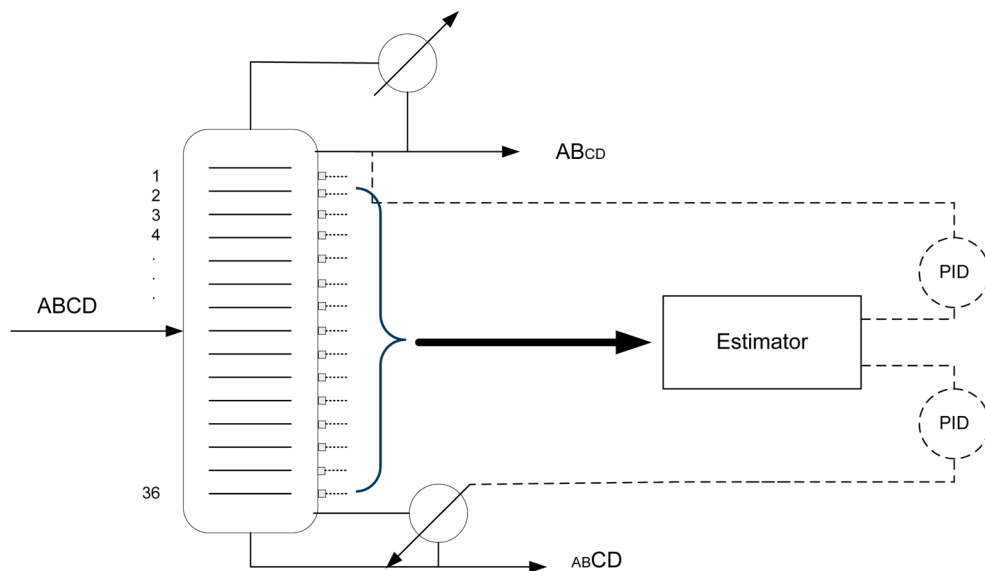


Figure 5. Schematic of the distillation process with estimator.

$$\mathbf{H}_{\text{PLS}} = \begin{bmatrix} -0.0045 & 0.0075 \\ 0.0113 & -0.0076 \\ -0.0036 & 0.0037 \\ -0.0038 & -0.0013 \\ 0.0074 & -0.0092 \\ -0.0055 & 0.0152 \\ -0.0022 & 0.0057 \\ 0.0011 & -0.0140 \end{bmatrix}$$

$$\mathbf{H}_{\text{CL}} = \begin{bmatrix} -0.0038 & 0.0080 \\ 0.0104 & -0.0082 \\ -0.0028 & 0.0041 \\ -0.0035 & -0.0011 \\ 0.0059 & -0.0101 \\ -0.0049 & 0.0156 \\ -0.0019 & 0.0059 \\ 0.0010 & -0.0141 \end{bmatrix}$$

Figure 6 shows the dynamic behavior of the model as disturbances happen, as well as the dynamic behavior 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 behavior can be corrected by feedback.

5.3. Further Examples. Some additional examples are provided by Skogestad et al.,¹⁴ where the new closed-loop estimator is compared with PCR and PLS. It is also suggested that the addition of “artificial noise” may provide additional degrees of freedom for our data-based method (\mathbf{B}_{CL}). This is particularly relevant when there are a large number of measurements (x), but relatively few samples, for example, for spectroscopic data, because it is then difficult to obtain a good estimate of \mathbf{W}_n^s , using eq 30. The idea is to add an extra diagonal matrix \mathbf{W}'_n to the end of $\mathbf{X}_{\text{opt}} = \tilde{\mathbf{F}}$, which contains the expected noise for each measurement x along its diagonal. This approach may have improved the results for estimator \mathbf{B}_{CL} in Figure 4(a)

6. DISCUSSION

6.1. Relationship to Self-Optimizing Control. This work originated from considering the “indirect control problem”,¹⁰ using the “exact local method” in self-optimizing control. In “indirect control”, the objective is to determine a set of controlled variables $c = \mathbf{H}x$, such that, by keeping c constant, we indirectly keep the primary variables y constant (or more specifically, at their desired set points y_s), despite disturbances d and measurement noise n^x . This can be viewed as a special case of “self-optimizing control” with cost function $J = \|y - \hat{y}\|_2$. 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 $c = \hat{y}$. This requires some explanation. In indirect control, we adjust the inputs u by feedback to keep $c = \mathbf{H}x = 0$ (constant). Note that we will generate the same inputs u (for a given d and n^x), also if we keep $c' = \mathbf{D}c = 0$, 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 is the estimate \hat{y} . However, if we look the other way around, then the optimal c will not necessarily correspond to an estimate of y (\hat{y}). 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 \hat{y} , which is, in fact, done when we select \mathbf{H} such that $\mathbf{H}\mathbf{G}_x = \mathbf{G}_y$, (see Theorem 4).

6.2. Comparison with the Work of Pannocchia and Brambilla.¹¹ Our paper provides an extension of the results of Pannocchia and Brambilla¹¹ on “steady-state closed-loop consistency” to include measurement noise as well. In addition, we have shown, in agreement with the results in the paper by Hori et al.,¹⁰ that we can always achieve “perfect consistency” for set-point 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,¹¹ 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 ($y =$

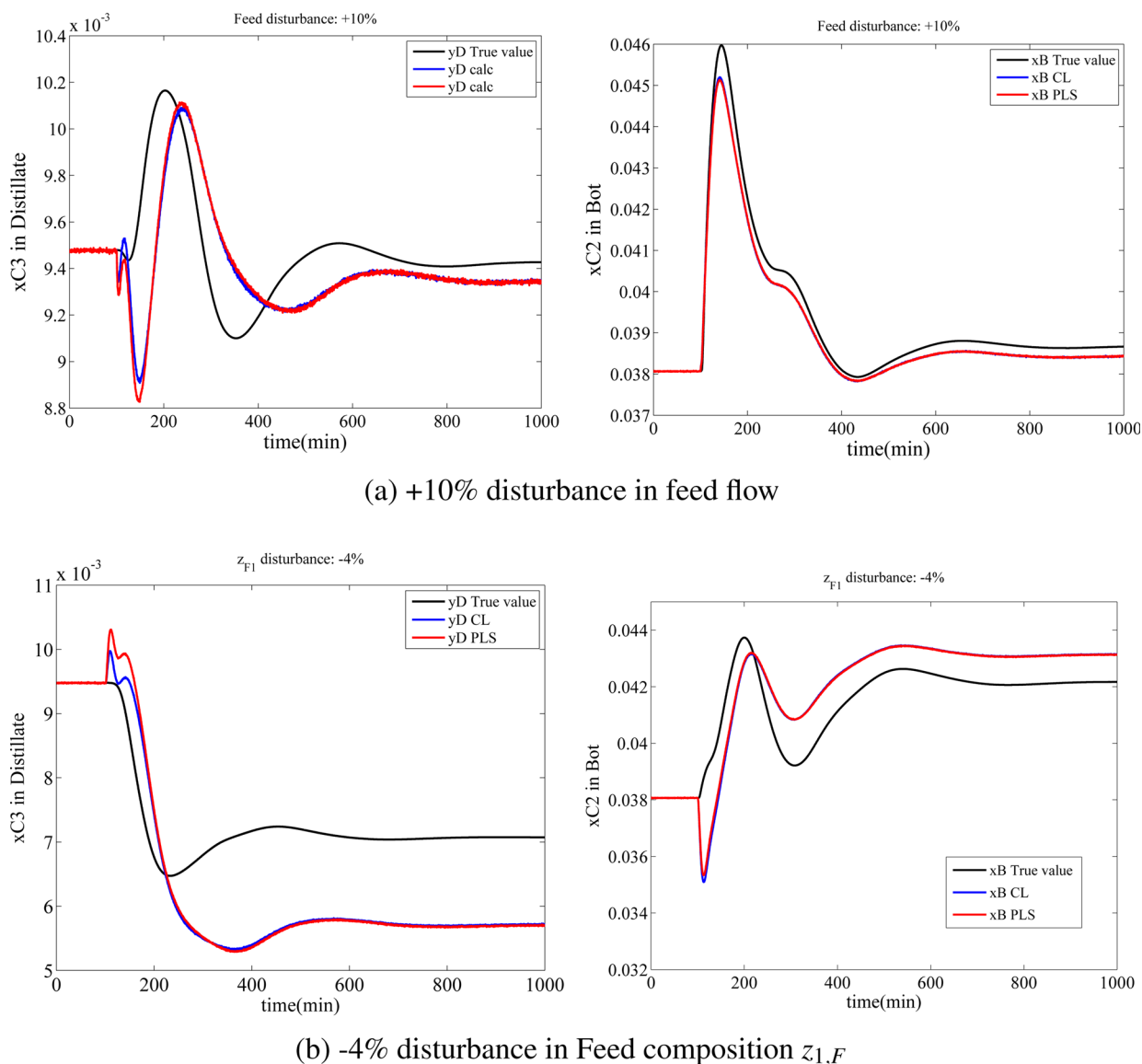


Figure 6. Estimated and model composition values for the case with two temperature controls and with the consideration of eight measurements.

$[x_D \ x_B]$), 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 column ends ($X = [T_D \ T_B]$), 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 $\sim 99.99\%$ water. At 1 atm, the boiling point of this mixture will be approximately $(0.9999 \times 100^\circ\text{C} + 0.0001 \times 65^\circ\text{C}) = 99.9965^\circ\text{C}$, whereas the boiling point of 100% water is 100.00°C . Thus, if we have a measurement error of more than 0.0035°C (which we certainly will have), then the temperature measurement will be useless to infer composition, since it would lead to predicting negative compositions. 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 include measurement noise.

6.3. Measurement Selection. The results presented in this paper also provide the basis for optimal measurement selection, which extends Algorithm 1 in Pannocchia and Brambilla¹¹ 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 measurements and 2 u measurements). 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.²¹

6.4. Comparison to Standard Data-Based Estimators. In least-squares (LS) regression, one gets $B = YX^{-1}$, or more generally $B = YX^\dagger$, where X^\dagger denotes the pseudo-inverse of X . In principal component regression (PCR), one uses $B = YX_a^\dagger$ where $X_a^\dagger = (\sum_{i=1}^a (1/\sigma_i)(v_i u_i^H))$ denotes the pseudo-inverse of $X = U\Sigma V^H$ with only a principal components included. Thus, in both LS and PCR, one inverts the X -matrix, while with the new estimation method, see eq 25 in Theorem 3, one considers only a part $X_{\text{opt}} = \tilde{F}$ of the transformed X -matrix. It is optimal because the corresponding Y data are zero. The proposed method seems

a bit similar to PLS in that we use the data for Y to affect the X -data (we get X_{opt} from X by using the SVD of Y).

Comparing the regression equations of our new estimator and PLS, we realize that the PLS method has one more degree of freedom (B_0), which provides an optimal centering of the data. We may include this degree of freedom into our method as follows. By assuming deviation variables, we may write (using here $H = B_{\text{CL}}$ to denote our data-based estimator)

$$Y - Y_0 = H(X - X_0) \quad (34)$$

or

$$Y = HX + H_0 \quad (35)$$

where $H_0 = Y_0 - HX_0$. By writing

$$H_0 = \text{diag}(H_0) \times 1\text{-vector} \quad (36)$$

Equation 35 then can be written as

$$Y' = H'X' \quad (37)$$

where $H' = [H \text{diag}(H_0)]$ and $X' = [X \text{1-vector}]$.

Thus, by just adding ones to the end of the X -data, one can optimize to determine H' , and then find H and H_0 . The general equation for B in PLS is²²

$$B_{\text{PLS}} = W_a(W_a^T X^T X W_a)^{-1} W_a^T w_1 \quad (38)$$

Comparing this with H for our closed-loop estimator in eq 25, we see that $X_{\text{opt}} = \tilde{F}$ is a variation of $W_a X$. X_{opt} is actually XV in our method, where V is the right singular vector. It acts as some form of W_a .

We must assume that $X_{\text{opt}} = \tilde{F}$ is full rank (invertible) to use the analytic expression in eq 25. If \tilde{F} does not have full rank, one may use some pseudo-inverse of \tilde{F} (similar to PCR). This adds degrees of freedom to the method, which, in PLS, is the size of the matrix W_a and is specified in the first step (the number of components in PLS). The problem of invertibility is solved by manipulating the W_a matrix.

The PLS method for univariate data is optimal in the prediction error sense.¹⁸ However, the PLS algorithm for multivariate data is not optimal in the same way as the PLS algorithm for univariate data. There are reports that, according to the literature, the PLS solutions using different approaches are not equivalent. For example, de Jong's SIMPLS solution²³ is not equivalent to Herman Wold's NIPALS solution.

As mentioned previously, reports from different studies have shown that the PLS method always gives a higher coefficient of determination (denoted as R^2 in statistics) than the PCR method (see Table 1 in the paper by Wentzell et al.⁷). However, some authors^{24–26} have taken a closer look on the shrinkage properties of PLS and have shown that PLS almost always can be improved in principle, so the regression method as such is not optimal.

7. CONCLUSION

In this paper, we have introduced a new class of static estimators based on minimizing the prediction error. We have considered four different cases, where the first three (S1–S3) correspond to cases where the estimator is used for monitoring, and the fourth case (S4) is when the estimator is used in closed-loop. The new estimators (Theorems 1–5) are derived based on the assumption that we have available a linear process model. If we only have data available, then these may directly be used for the three monitoring estimators (S1–S3). For the closed-loop estimator (S4), we have proposed a method to extract the data (see eqs 29 and 30). For the data-based case, the new estimators

have been compared with the established PCR and PLS estimators, and the results were found to be comparable (see Figure 4). For a specific case, our new estimators should be better, because they are optimal in terms of minimizing the prediction error, but PCR and PLS are found to generally give good predictions.

AUTHOR INFORMATION

Corresponding Author

*E-mail: skoge@nt.ntnu.no.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

David Di Ruscio and Inge Svein Helland are acknowledged for discussions about partial least-squares (PLS), and Ivar J. Halvorsen is acknowledged for discussions.

REFERENCES

- (1) Mejdell, T.; Skogestad, S. Estimation of distillation compositions from multiple temperature measurements using partial-least-squares regression. *Ind. Eng. Chem. Res.* **1991**, *30*, 2543–2555.
- (2) Zhang, J. Inferential Feedback Control of Distillation Composition Based on PCR and PLS Models. *Proc. Am. Control Conf.* **2001**, *2001*, 1196–1201.
- (3) Bhartiya, S.; Whiteley, R. J. Development of inferential measurements using neural networks. *ISA Trans.* **2001**, *40*, 307–323.
- (4) Yan, W.; Shao, H.; Wang, X. Soft sensing modeling based on support vector machine and Bayesian model selection. *Comput. Chem. Eng.* **2004**, *28*, 1489–1498.
- (5) Massy, W. F. Principal Component Regression in Exploratory Statistical Research. *J. Am. Stat. Assoc.* **1965**, *60*, 234–246.
- (6) Wold, H. PLS path models with latent variables: the nipals approach. In *Quantitative Sociology: International Perspectives on Mathematical and Statistical Modeling*; Blalock, H. M., Aganbegian, A., Borodkin, F. M., Boudon, R., Cappecchi, V., Eds.; Academic Press: New York, 1975; pp 307–353.
- (7) Wentzell, P. D.; Montoto, V. Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures. *Chemom. Intell. Lab. Syst.* **2003**, *65*, 257–279.
- (8) Joseph, B.; Brosilow, C. B. Inferential Control of Processes: Part I. Steady State Analysis and Design. *AIChE J.* **1978**, *24*, 485–492.
- (9) Kresta, J. V.; Marlin, T. E.; MacGregor, J. F. Development of Inferential Process Models using PLS. *Comput. Chem. Eng.* **1994**, *18*, 597–611.
- (10) Hori, E. S.; Skogestad, S.; Alstad, V. Perfect Steady-State Indirect Control. *Ind. Eng. Chem. Res.* **2005**, *44*, 863–867.
- (11) Pannocchia, G.; Brambilla, A. How auxiliary variables and plant data collection affect closed-loop performance of inferential control. *J. Process Control* **2007**, *17*, 653–663.
- (12) Tronci, S.; Bezzo, F.; Barolo, M.; Baratti, R. Geometric observer for a distillation column: Development and experimental testing. *Ind. Eng. Chem. Res.* **2005**, *44*, 9884–9893.
- (13) Alstad, V.; Skogestad, S.; Hori, E. S. Optimal measurement combinations as controlled variables. *J. Process Control* **2009**, *19*, 138–148.
- (14) Skogestad, S.; Yelchuru, R.; Jäschke, J. 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*; Huba, M. et al., Eds.; STU Bratislava and NTNU: Trondheim, Norway, 2011; pp 53–86 (ISBN 978 80 968627 3 3).
- (15) Rosipal, R.; Krämer, N. Overview and Recent Advances in Partial Least Squares. In *Subspace, Latent Structure and Feature Selection*; Saunders, C., Grobelnik, M., Gunn, S., Shawe-Taylor, J., Eds.; Lecture

Notes in Computer Science, Vol. 3940; Springer-Verlag: Berlin, Heidelberg, 2006; pp 34–51.

(16) Wold, H. Soft Modelling by Latent Variables; the Nonlinear Iterative Partial Least Squares Approach. In *Perspectives in Probability and Statistics, Papers in Honour of M. S. Barlett*; Gani, J., Ed.; Academic Press: London, 1975; pp 117–142.

(17) Ergon, R. Noise Handling Capabilities of Multivariate Calibration Methods. *Model. Identif. Control* **2002**, *23*, 259–273.

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

(19) Elden, L. Partial Least-squares vs. Lanczos Bidiagonalization—I: Analysis of a Projection Method for Multiple Regression. *Comput. Stat. Data Anal.* **2004**, *46*, 11–31.

(20) Skogestad, S. Dynamics and Control of Distillation Columns—A Critical Survey. *Model. Identif. Control* **1997**, *18*, 177–217.

(21) Kariwala, V.; Cao, Y. Bidirectional branch and bound for controlled variable selection. Part II: Exact local method for self-optimizing control. *Comput. Chem. Eng.* **2009**, *33*, 1402–1412.

(22) Di Ruscio, D. The Partial Least Squares Algorithm: A Truncated Cayley-Hamilton Series Approximation Used to Solve the Regression Problem. *Model. Identif. Control* **1998**, *19*, 117–140.

(23) de Jong, S. SIMPLS: An alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 251–263.

(24) Butler, N. A.; Denham, M. C. The Peculiar Shrinkage Properties of Partial Least Squares Regression. *J. R. Stat. Soc.* **2000**, *B 62*, 585–593.

(25) Lingjærde, O. C.; Christophersen, N. Shrinkage Structure of Partial Least Squares. *Scand. J. Stat.* **2000**, *27*, 459–473.

(26) Helland, I. S. Some Theoretical Aspects of Partial Squares Regression. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 97–107.