Computers and Chemical Engineering 179 (2023) 108429

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/cace

An industrial case study on the combined identification and offset-free control of a chemical process☆

Steven J. Kuntz[a,*], James J. Downs[b], Stephen M. Miller[b], James B. Rawlings[a]

[a] Department of Chemical Engineering, University of California, Santa Barbara, CA 93106, United States
[b] Eastman Chemical Company, Kingsport, TN 37662, United States

ARTICLE INFO

Keywords:
System identification
Identification for control
Model predictive control
Subspace identification
Closed-loop identification
Disturbance identification

ABSTRACT

For three decades, model predictive control (MPC) has been the flagship advanced control method in the chemical process industries. However, most implementations still use heuristic methods for designing MPC estimators, especially for offset-free MPC implementations. In this paper, we present a recently developed maximum likelihood-based method for the identification of linear augmented disturbance models for use in offset-free MPC. This method provides noise covariances that are used to derive Kalman filters and moving horizon estimators, forgoing the need for manual design and tuning of the estimator. The method is extended to handle closed-loop plant data. The proposed identification method and estimator design are evaluated in industrial-scale, real-world case study of a process at Eastman Chemical's Kingsport plant. Using this identified model, we reduced the mean stage cost by 38% compared to the performance of the existing, hand-tuned MPC model.

1. Introduction

Model predictive control (MPC) is widely used in the chemical process industries as an advanced feedback control method (Qin and Badgwell, 2003). Some important factors in the success of MPC are its inherent robustness to disturbances and plant-model mismatch, and the ability to track setpoints without offset (Rawlings et al., 2020, pp. 46–59, 204–214). As is often noted by industrial practitioners, MPC can be quite forgiving with respect to model errors, aging of the plant, changes in environmental conditions, and changes in operating conditions. Practitioners have long used heuristic or out-of-date models, without rigorous methods of identifying both plant and disturbance models (Lee and Yu, 1994; Caveness and Downs, 2005). Despite this, the performance and lifetime of an MPC deployment is tightly connected to the quality of the model over time (Canney, 2003; Darby and Nikolaou, 2012). As stake holders continue to demand greater and more consistent performance from their processes, they require a system of best practices for identifying plant and disturbance models.

Traditionally, MPC implementations have relied on linear finite impulse response (FIR) plant models (Qin and Badgwell, 2003; Darby and Nikolaou, 2012) with which a dynamic matrix control (DMC) (Cutler and Ramaker, 1980) or Identification and Command (IDCOM) (Richalet et al., 1978) algorithm is implemented. A few products, such as the Shell Multivariable Optimizing Controller (SMOC) (Marquis and Broustail, 1988; Yousfi and Tournier, 1991) and Adersa's predictive functional control (PFC) algorithm, rely solely on a linear state-space plant model. Darby and Nikolaou (2012) note that recent MPC products have shifted away from FIR models and towards linear state-space models. This shift is motivated by a number of shortcomings of the FIR approach, most notably: (1) the inability to handle unstable and integrating systems without modification, (2) the overparameterization of the underlying linear system (especially for slow processes), (3) the difficulty of formulating estimators, and (4) the fact that FIR models are a special case of the linear state-space model (Lee et al., 1994; Lundström et al., 1995).

Other plant model formulations include autoregressive models (e.g., ARMA and CARIMA models) (Clarke et al., 1987a,b; Clarke, 1991; Sun et al., 2011) and transfer function models (Ljung, 1999). Both model types require complicated estimator formulations and their identification algorithms are typically formulated for single-input single-output (SISO) systems. As such, multi-input multi-output (MIMO) models are typically constructed from individually fit SISO models. Transfer function models must be realized as state-space models in order to formulate controller constraints. As with FIR models, every autoregressive and transfer function model can be realized as a state-space model (Ho and Kalman, 1966; Akaike, 1974).

☆ This work is an expanded version of a conference paper presented at FOCAPO/CPC 2023 (Kuntz et al., 2023).

* Corresponding author.
E-mail addresses: skuntz@ucsb.edu (S.J. Kuntz), jbraw@ucsb.edu (J.B. Rawlings).

https://doi.org/10.1016/j.compchemeng.2023.108429
Received 10 May 2023; Received in revised form 8 September 2023; Accepted 13 September 2023
Available online 18 September 2023

To identify the plant model, practitioners typically fit a linear model to step response data, although it is also possible to linearize a physics-based plant model (Caveness and Downs, 2005; Rawlings et al., 2020). Neither approach provides the noise covariance estimates required to design an estimator for MPC implementation. While subspace methods—such as canonical variate analysis (CVA) (Larimore, 1983), N4SID (Van Overschee and De Moor, 1994), or MOESP (Verhaegen, 1994)—can be used to identify estimate the process and measurement noise covariances, these methods can only identify controllable and observable realizations (Qin, 2006), and the disturbance model contains uncontrollable integrating modes (Muske and Badgwell, 2002; Pannocchia and Rawlings, 2003). Disturbance models may be tuned under strong assumptions on the process and measurement noises (Lee et al., 1994; Lee and Yu, 1994), but the required assumptions are not general, producing suboptimal estimator performance. Autocovariance least squares (ALS) can identify the complete disturbance model, but it does not identify the plant model (Odelson et al., 2006). Additionally, there is a trade-off between the computational complexity of ALS and the variance of the ALS estimates because the optimal least squares weighting matrix is a function of the covariances to be estimated (Rajamani and Rawlings, 2009; Zagrobelny and Rawlings, 2015; Arnold and Rawlings, 2022). Kuntz and Rawlings (2022) presented the first identification algorithm that provides estimates of both the state-space model coefficients and the disturbance noise covariance required to implement an offset-free MPC.

Most of the MPC deployment cost is incurred during plant identification due to the commonality of open-loop identification experiments, where product quality is difficult if not impossible to maintain, and the process must be perturbed from the optimal operating point in order to acquire quality data (Canney, 2003; Zhu, 2006). As a result, closed-loop identification experiments are an opportunity for significant safety and profitability improvements in chemical process control. Closed-loop identification experiments can then be conducted online, at and around the optimal operating point, negating the cost of opening the loop to perform the experiment. New MPCs can be implemented on processes controlled with other methods (PID, DMC, etc.) and existing MPCs be significantly improved with re-identified models. Closed-loop experiments can be conducted via setpoint perturbations that are more predictable and reliable than open-loop input perturbations. Moreover, the control loop is never broken, so the MPC is always enforcing constraints throughout the experiment.

Canney (2003) points out that MPC performance decays over time after deployment, and proposes MPC upkeep be a continuous process of algorithm improvement, where the model, MPC tuning, and organizational details are adjusted as necessary. A closed-loop disturbance model identification method can be applied to continuous offset-free MPC monitoring and upkeep. Previous attempts at continuous MPC monitoring and upkeep simply attempt to detect (and sometimes diagnose the source of) plant-model mismatch (Harrison and Qin, 2009; Pannocchia and De Luca, 2012; Kheradmandi and Mhaskar, 2018). However these algorithms rely on heuristic cutoffs for the alarm thresholds because they are based on LTI system order estimation. With the full set of parameter estimates, there is a future possibility of advanced offset-free MPC monitoring schemes with rigorous performance guarantees.

Closed-loop experimentation requires an existing controller, meaning open-loop experiments for MPC design or PID tuning are still necessary. To this end, we suggest suboptimal but safe experiments be done using traditional step-response designs, or loops be initially closed with PID methods. While the algorithm proposed herein and in Kuntz and Rawlings (2022) will still handle open-loop step responses. At a later date, a closed-loop identification experiment may be run to refine and re-identify the model. The only advantage of open-loop methods, such that the one in Kuntz and Rawlings (2022), are their relative simplicity compared to closed-loop methods, such as the one proposed in this paper.

In this paper, we present a closed-loop extension of the algorithm proposed in Kuntz and Rawlings (2022) and demonstrate its efficacy in an industrial case study. Our method systematizes the identification of new offset-free MPC models and design of new MPC estimators, allowing practitioners save time and achieve optimal estimator performance. To do this, we combine the plant modeling and disturbance modeling steps by passing information about state estimates between identification steps. Because state information is passed between steps, each step can be formulated as a linear regression problem for which closed-form solutions are readily available (Rao, 1973; Anderson, 2003). The plant modeling step is a regularized version of the closed-loop identification procedure outlined by Larimore (1983, 1997, 2005). To validate the viability of our method in the wider chemical process industries, we performed a case study on an existing process at Eastman Chemical's Kingsport, Tennessee location. The newly identified model shows clear improvement from the older step-response model, and the closed-loop performance is improved as measured by the controlled variable tracking error. Moreover, we used a closed-loop experimental design that is desirable to operations engineers for its simplicity, safety, and ability to produce predictably high-quality data. The case study serves as a template for using this new method to improve existing MPC performance.

In Section 2 we define the plant and models, present the offset-free MPC algorithm, and discuss some MPC properties that motivate the identification algorithm. In Section 3 we describe the closed-loop subspace identification procedure used in the case study. In Section 4 we describe the disturbance model identification method used in the case study. In Section 5, we present our case study of the combined plant and disturbance identification method as applied to a reactor at Eastman Chemical Company's Kingsport plant. Finally, in Section 6, we summarize the methods and case study, and discuss future work.

Notation. The set of real numbers, real n-vectors, and real $m \times n$ matrices are denoted $\mathbb{R}$, $\mathbb{R}^n$, and $\mathbb{R}^{m\times n}$, respectively. The set of integers, nonnegative integers, and integers from m to $n \geq m$ (inclusive) are denoted $\mathbb{I}$, $\mathbb{I}_{\geq 0}$, and $\mathbb{I}_{m:n}$, respectively. We denote by I_n and $0_{m\times n}$ the $n \times n$ identity matrix and $m \times n$ zero matrix, respectively. Subscripts are omitted when the dimensions are clear from context. The transpose and pseudoinverse of $A \in \mathbb{R}^{m\times n}$ are denoted A' and $A^\dagger$, respectively. The inverse of $A \in \mathbb{R}^{n\times n}$, if it exists, is denoted A^{-1}. The trace and determinant of $A \in \mathbb{R}^{n\times n}$ are denoted $\mathrm{tr}(A)$ and $|A|$, respectively. The Krocker product of $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$ is denoted $A \otimes B$. We denote that $W \in \mathbb{R}^{n\times n}$ is positive (semi)definite by $W > 0$ ($W \geq 0$). We denote the positive semidefinite square root of $W \geq 0$ as $W^{1/2} \geq 0$, where $W = (W^{1/2})^2$. We denote the 2-norm as $\|\cdot\|$ for both vector and matrix arguments. For any positive (semi)definite matrix $W \in \mathbb{R}^{n\times n}$, we define the W-(semi)norm, denoted $\|\cdot\|_W$, as $\|x\|_W = \sqrt{x^\top W x}$ for all $x \in \mathbb{R}^n$.

For any signal $(a(k))_{k\in\mathbb{I}_{\geq 0}}$, we use the shorthand $a^+ = a(k+1)$ and denote the length-n past and future horizons as

$$A_{-n}(k) := \begin{bmatrix} a(k-1) \\ \vdots \\ a(k-n) \end{bmatrix}, \quad A_n(k) := \begin{bmatrix} a(k) \\ \vdots \\ a(k+n-1) \end{bmatrix}$$

For any two signals $(a(k))_{k\in\mathbb{I}_a}$ and $(b(k))_{k\in\mathbb{I}_b}$, we denote the sample covariance operator as $S(a,b) = \frac{1}{N_{ab}} \sum_{k\in\mathbb{I}_a\cap\mathbb{I}_b} a(k)b(k)'$ where N_{ab} is the number of elements in $\mathbb{I}_a \cap \mathbb{I}_b$, and the index sets $\mathbb{I}_a, \mathbb{I}_b \subseteq \mathbb{I}$ are implied from context.

We denote that a random vector x has a distribution $\mathcal{D}$ by $x \sim \mathcal{D}$, and that the stochastic process $x(k)$ is independently and identically distributed as $\mathcal{D}$ by $x(k) \overset{iid}{\sim} \mathcal{D}$ or $x \overset{iid}{\sim} \mathcal{D}$. We denote that a random vector x has a Gaussian distribution with mean μ and covariance Σ by $x \sim \mathcal{N}(\mu, \Sigma)$. We denote the probability density function of x (conditioned on y) as $p(x)$ ($p(x|y)$).

2. Problem statement

In this section, we describe the plant and linear models common to offset-free MPC theory and practice. We also describe the offset-free MPC algorithm that is used in the case study and discuss some theoretical results that motivate some choices in the identification algorithm.

2.1. Systems of interest

We are concerned with offset-free control of the following discrete-time plant,

$$x_p^+ = f_p(x_p, u, w_p) \tag{1a}$$

$$y = h_p(x_p, v_p) \tag{1b}$$

where $x_p \in \mathbb{R}^{n_p}$ is the plant state, $u \in \mathbb{R}^{n_u}$ is the input, $y \in \mathbb{R}^{n_y}$ is the measurement, and $w_p \in \mathbb{R}^{n_w}, v_p \in \mathbb{R}^{n_v}$ are the plant process and measurement disturbances, respectively. Standard MPC relies on a state-space model to formulate the estimator and optimal control problem,

$$x^+ = Ax + Bu + w \tag{2a}$$

$$y = Cx + v \tag{2b}$$

$$\begin{bmatrix} w \\ v \end{bmatrix} \overset{iid}{\sim} \mathcal{N}\left(0, \begin{bmatrix} Q_w & \\ & R_v \end{bmatrix}\right) \tag{2c}$$

where $x \in \mathbb{R}^n$ is the state and $w \in \mathbb{R}^n, v \in \mathbb{R}^{n_y}$ are the process and measurement noises, respectively. We assume (w, v) is uncorrelated in time. In the presence of persistent disturbances or errors, offset-free tracking is achieved by augmenting the standard state-space model with uncontrollable integrating modes, called the augmented disturbance model (Muske and Badgwell, 2002; Pannocchia and Rawlings, 2003),

$$x^+ = Ax + Bu + B_d d + w \tag{3a}$$

$$d^+ = d + w_d \tag{3b}$$

$$y = Cx + C_d d + v \tag{3c}$$

$$\begin{bmatrix} w \\ w_d \\ v \end{bmatrix} \overset{iid}{\sim} \mathcal{N}(0, S_d) \tag{3d}$$

where $d \in \mathbb{R}^{n_d}$ is the disturbance and $w_d \in \mathbb{R}^{n_d}$ is the disturbance driving noise. Again, we assume (w, w_d, v) is uncorrelated in time. The goal of the identification algorithm is to estimate the parameters of the augmented disturbance model (3) from only input–output data (u, y).

Remark 1. The models (2), (3) differ from general linear time-invariant (LTI) systems in that they both lack passthrough terms Du, and the model (2) lacks a cross-covariance $\mathrm{cov}(w, v)$, i.e., they are special cases of

$$\begin{aligned} x^+ &= Ax + Bu + w \\ y &= Cx + Du + v \end{aligned} \qquad \begin{bmatrix} w \\ v \end{bmatrix} \overset{iid}{\sim} \mathcal{N}(0, S)$$

Passthrough terms are not included in our MPC formulation, so we do not include them in the model. The restriction that $\mathrm{cov}(w, v) = 0$ in the model (2) facilitates estimation under the restriction $D = 0$. Estimating a model with passthrough $D \neq 0$ is a standard generalization of the methods discussed herein.

2.2. Offset-free model predictive control

Offset-free MPC consists of three distinct problems that are solved at each time step: estimation, target calculation, and regulation. The goal is firstly to remove offset in the *controlled variables* $r(k) = Hy(k) \in \mathbb{R}^{n_r}$ and secondly to minimize the distance from a pair of input–output setpoints $(u_{\mathrm{sp}}, y_{\mathrm{sp}}) \in \mathbb{R}^{n_u+n_y}$. This case study uses the steady-state Kalman filter for estimation, uses an infinite horizon optimal control problem for regulation, includes a steady-state target problem, and incorporates box constraints on the inputs and outputs.[1]

State and disturbance estimator. For stochastic LTI systems of the forms (2), (3), the Kalman filter is the optimal state estimator. For the augmented disturbance model (3), the steady-state Kalman filter takes the following form,

$$\begin{bmatrix} \hat{x} \\ \hat{d} \end{bmatrix}^+ = \begin{bmatrix} A & B_d \\ 0 & I \end{bmatrix} \begin{bmatrix} \hat{x} \\ \hat{d} \end{bmatrix} + \begin{bmatrix} B \\ 0 \end{bmatrix} u + \begin{bmatrix} L_x \\ L_d \end{bmatrix} \left(y - \begin{bmatrix} C & C_d \end{bmatrix} \begin{bmatrix} \hat{x} \\ \hat{d} \end{bmatrix} \right) \tag{4}$$

where $L := \begin{bmatrix} L_x' & L_d' \end{bmatrix}'$ is the steady-state Kalman filter gain for the augmented disturbance model (3). We refer the reader to Kwakernaak and Sivan (1972) and Hespanha (2018) for a classical treatment of the linear optimal estimation problem and to Rawlings et al. (2020, pp. 27–46) for a derivation of the optimal filter gain $L := \begin{bmatrix} L_x' & L_d' \end{bmatrix}'$ from least squares theory.

Steady-state target problem. Given the current filtered disturbance estimate $\hat{d}(k)$, the steady-state targets $(x_s(k), u_s(k), y_s(k))$ defined as the solutions to the following steady-state target problem (SSTP).

$$\min_{x,u,y} \frac{1}{2}\|y - y_{\mathrm{sp}}(k)\|_{Q_s}^2 + \frac{1}{2}\|u - u_{\mathrm{sp}}(k)\|_{R_s}^2 \tag{5a}$$

$$\text{s.t. } x = Ax + Bu + B_d \hat{d}(k), \tag{5b}$$

$$\underline{y} \le y = Cx + C_d \hat{d}(k) \le \overline{y}, \tag{5c}$$

$$\underline{u} \le u \le \overline{u}, \tag{5d}$$

$$r_{\mathrm{sp}}(k) = Hy \tag{5e}$$

where $(\underline{u}, \overline{u})$ are the input bounds, $(\underline{y}, \overline{y})$ are the output bounds, (Q_s, R_s) are positive semidefinite weighting matrices. Notice that the steady-state targets are functions of only the current disturbance estimate $\hat{d}(k)$ and the current setpoints $(u_{\mathrm{sp}}(k), y_{\mathrm{sp}}(k), r_{\mathrm{sp}}(k))$.

Infinite horizon optimal control problem. The regulator is defined as an infinite horizon optimal control problem that is solved about the steady-state targets $(x_s(k), u_s(k), y_s(k))$. The control law is defined as

$$u(k) = \tilde{u}_0^*(k) + u_s(k) \tag{6a}$$

where the $\tilde{x}_i^*(k)$ and $\tilde{u}_i^*(k)$ denote solutions to

$$\min_{\substack{\tilde{x}_0, \tilde{x}_1, \ldots \\ \tilde{u}_0, \tilde{u}_1, \ldots}} \frac{1}{2} \sum_{i=0}^{\infty} \|C\tilde{x}_i\|_Q^2 + \|\tilde{u}_i\|_R^2 + \|u_i - u_{i-1}\|_M^2 \tag{6b}$$

$$\text{s.t. } \tilde{x}_0 = \hat{x}(k) - x_s(k), \tag{6c}$$

$$\tilde{x}_{i+1} = A\tilde{x}_i + B\tilde{u}_i, \tag{6d}$$

$$\underline{y} \le C\tilde{x}_i + y_s(k) \le \overline{y}, \tag{6e}$$

$$\underline{u} \le \tilde{u}_i + u_s(k) \le \overline{u} \tag{6f}$$

where (Q, R, M) are positive semidefinite weighting matrices.[2]

2.3. Offset-free sufficient conditions

Muske and Badgwell (2002) first established sufficient conditions under which the offset-free MPC (4)–(6) with a separable disturbance model,

$$B_d = \begin{bmatrix} \overline{B}_d & 0 \end{bmatrix}, \qquad C_d = \begin{bmatrix} 0 & \overline{C}_d \end{bmatrix}$$

[1] The output constraints are implemented as soft constraints in the optimizer.

[2] The infinite-horizon optimal control problem (6) is practically solved as a finite-horizon optimal control problem, where the horizon length is taken sufficiently large to approximate the infinite-horizon controller.

applied to a linear plant converges to the controlled variable setpoints r_{sp}. This was generalized to linear models of the form (3) by Pannocchia and Rawlings (2003). Finally, Morari and Maeder (2012) generalized the conditions to nonlinear plants and models. We restate the offset-free conditions for linear models in the following theorem.

Theorem 1 (*Pannocchia and Rawlings, 2003*). *Consider a system controlled by the offset-free MPC (4)–(6). Assume that*

1. *the disturbance state is of the same dimension as the measurement ($n_d = n_y$) and*
2. *the augmented disturbance model (3) is detectable.*

If the closed-loop system is stable and the constraints are not active at steady state, then there is zero offset in the controlled variables at steady state,

$$\lim_{k\to\infty} Hy(k) = r_{sp}$$

Remark 2. Despite the fact that Theorem 1 does not explicitly mention control of nonlinear plants, the results are widely applicable to both linear and nonlinear plants, without disturbances and with asymptotically constant disturbances. This is because Theorem 1 does not make statements about controller stability, but simply states sufficient conditions for which a stable controller also has zero offset. Pannocchia and Rawlings (2003) demonstrate the validity of Theorem 1 in the control of a non-isothermal reactor model.

An immediate consequence of Theorem 1 is that, to achieve offset-free control with offset-free MPC, it is important to have a detectable model. To this end, we have the following result.

Lemma 2 (*Pannocchia and Rawlings, 2003*). *The augmented disturbance model (3) is detectable if and only if the standard model (2) is detectable and*

$$\operatorname{rank}\begin{bmatrix} A - I_n & B_d \\ C & C_d \end{bmatrix} = n + n_d \tag{7}$$

The so-called *offset-free rank condition* (7) is important in formulating disturbance models for the offset-free MPC algorithm. One can replace the third condition of Theorem 1 with the rank condition (7). It turns out that, in the same way that a state-space realization is only unique up to a similarity transformation, any *detectable* disturbance model is only unique up to a similarity transformation. In fact, the Kalman filter behavior is equivalent under this similarity transformation, so if disturbances are “misassigned” in the model there is no effect on the closed-loop system.

Lemma 3 (*Rajamani et al., 2009*). *Consider the augmented system*

$$x^+ = Ax + Bu + \tilde{B}_d\tilde{d} + w \tag{8a}$$

$$\tilde{d}^+ = \tilde{d} + \tilde{w}_d \tag{8b}$$

$$y = Cx + \tilde{C}_d\tilde{d} + v \tag{8c}$$

$$\begin{bmatrix} w \\ \tilde{w}_d \\ v \end{bmatrix} \overset{iid}{\sim} \mathcal{N}(0, \tilde{S}_d) \tag{8d}$$

If the standard model (2) is detectable, then the augmented disturbance models (3), (8) are detectable if and only if both satisfy the offset-free rank condition (7). Moreover, there exists a choice of $\tilde{S}_d$ such that the models (3), (8) have equivalent Kalman filter innovations.

The consequence of Lemma 3 is that, given a standard model (2), one can “design” the disturbance model to be maximally interpretable, so long as it satisfies the rank condition (7). Typical “designs” are the *output disturbance model* $(B_d, C_d) = (0, I)$ and the *input disturbance model* $(B_d, C_d) = (B, 0)$.

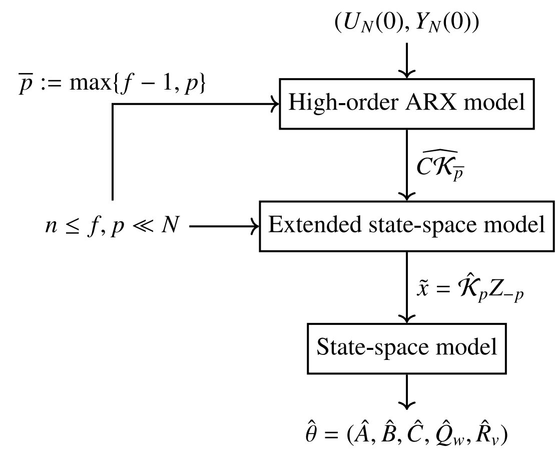

Fig. 1. Outline of the closed-loop subspace method, based on the work of Larimore (1983, 1997, 2005), used in this case study.

3. Closed-loop subspace identification

In this section, we describe a modification of the CVA algorithm of Larimore (1983, 1997, 2005). The algorithm’s goal is to estimate the parameters $\theta = (A, B, C, Q_w, R_v)$ of the model (2) from input–output data $(U_N(0), Y_N(0))$. The algorithm, outlined in Fig. 1, can be viewed as a nested modeling procedure using maximum likelihood (ML) at each step to compute parameter estimates. We refer the reader to Gong and Samaniego (1981) for a theoretical justification of nested ML estimation. The algorithm takes two basic steps. First, we determine a state sequence $(\hat{x}(k))_{k\in\mathbb{I}_{p:N}}$ via approximations of the ML problem corresponding to the *marginal density* in $Y_N(0)$,

$$\max_{\theta} L_N(\theta) := \ln p(Y_N(0)|U_N(0), \theta) \tag{9}$$

where $\theta = (A, B, C, Q_w, R_v)$ are the system parameters and $p \ge n$ is an integer to be defined. Then, we solve the ML problem corresponding to the *joint density* in $(X_{N_s+1}(p), Y_{N_s}(p))$,

$$\max_{\theta} L_N^{\text{SS}}(X_{N_s+1}(p), \theta) := \ln p(X_{N_s+1}(p), Y_{N_s}(p)|U_{N_s}(p), \theta) \tag{10}$$

using the previously estimated state sequence in place of $X_{N_s+1}(p)$, where $N_s := N - p$.

The core model of the algorithm is the model (2), which, under the assumption that some approximate state sequence $(\tilde{x}(k))$ is available, can be used to solve for parameter estimates $\hat{\theta}$ via multivariate or regularized regression methods. However, we must first construct state estimates $(\tilde{x}(k))$.

In lieu of the states, we can use the Kalman filter state estimates, which are represented by past input–output data. There exists a steady-state Kalman gain K and innovation error covariance R_e such that $A_K := A - KC$ is stable and

$$\hat{x}^+ = A_K\hat{x} + B_K z \tag{11a}$$

$$e := y - C\hat{x} \overset{iid}{\sim} \mathcal{N}(0, R_e) \tag{11b}$$

where $B_K := \begin{bmatrix} B & K \end{bmatrix}$, $\hat{x} \in \mathbb{R}^n$ are the state estimates, and $z := \begin{bmatrix} u' & y' \end{bmatrix}'$ is the combined input–output data (Kwakernaak and Sivan, 1972; Hespanha, 2018). Given any $n \le p \ll N$ chosen large enough so that $A_K^p \approx 0$, we can recursively solve the Kalman predictor (11) to write the state as follows,

$$\hat{x}(k) = A_K^p\hat{x}(k-p) + \mathcal{K}_p Z_{-p}(k) \approx \mathcal{K}_p Z_{-p}(k)$$

where $\mathcal{K}_p := \begin{bmatrix} B_K & A_K B_K & \dots & A_K^{p-1}B_K \end{bmatrix}$. Therefore, estimating $\mathcal{K}_p$ also provides state estimates $\tilde{x} := \mathcal{K}_p Z_{-p}$ for estimation of the parameters θ in the model (2).

To estimate the controllability matrix $\mathcal{K}_p$, we rewrite the Kalman predictor (11) into the following extended state-space model,

$$\begin{aligned} Y_f(k) &= \mathcal{O}_f A_K^p \hat{x}(k-p) + \mathcal{H}_{f,p} Z_{-p}(k) + \mathcal{G}_f Z_f(k) + E_f(k) \\ &\approx \mathcal{H}_{f,p} Z_{-p}(k) + \mathcal{G}_f Z_f(k) + E_f(k) \end{aligned} \tag{12}$$

where $n \leq f \ll N$ is a user-provided integer,

$$\mathcal{O}_f := \begin{bmatrix} C \\ CA_K \\ \vdots \\ CA_K^{f-1} \end{bmatrix}, \quad \mathcal{G}_f := \begin{bmatrix} 0 & & & \\ G_1 & 0 & & \\ \vdots & \ddots & \ddots & \\ G_{f-1} & \dots & G_1 & 0 \end{bmatrix},$$

$G_i := CA_K^{i-1} B_K$ are the Markov parameters of the Kalman predictor (11), and $\mathcal{H}_{f,p} := \mathcal{O}_f \mathcal{K}_p$ is a block-Hankel matrix of the Markov parameters. Notice that, if the model (2) is minimal,[3] the coefficient matrix $\mathcal{H}_{f,p}$ has rank n. Assuming we have access to some Markov parameter estimates $\hat{G}_i$, the extended state-space model (12) takes the form of a classic rank-reduced regression problem for which closed-form solutions are well-known (Larimore, 1983; Anderson, 1999).[4] However, we must first obtain Markov parameter estimates $\hat{G}_i$.

Jansson (2003) first proposed "pre-estimation" of the Markov parameters G_i from the following ARX model,

$$\begin{aligned} y(k) &= CA_K^p \hat{x}(k-p) + C\mathcal{K}_{\bar{p}} Z_{-\bar{p}}(k) + e(k) \\ &\approx C\mathcal{K}_{\bar{p}} Z_{-\bar{p}}(k) + e(k) \end{aligned} \tag{13}$$

where $\bar{p} := \max\{f-1, p\}$, and the coefficient matrix contains the first $\bar{p}$ Markov parameters, $C\mathcal{K}_{\bar{p}} = \begin{bmatrix} G_1 & G_2 & \dots & G_{\bar{p}} \end{bmatrix}$. Noting that the ARX model (13) takes the form of a classical multivariate regression problem, we can estimate G_i, in closed-form, using multivariate regression methods.

For brevity, we defer the derivation of ML estimators of the models (2), (12), (13) to Appendix. Other classic subspace algorithms—such as N4SID (Van Overschee and De Moor, 1994) and MOESP (Verhaegen, 1994)—can be used to supply parameters to the disturbance model identification method of Section 4, so long as a Markov parameter "pre-estimation" step is included. Van Overschee and De Moor (1995) showed that the classic subspace algorithms (CVA, N4SID, and MOESP) are equivalent up to formulation of the estimation objective for estimation of the model (12).[5] Only the method of Larimore (1983, 1997, 2005) uses ML estimation at each step of the algorithm, making it the logical choice for integration with the ML-based disturbance model identification. To use the method of Section 4, one should take care to use methods that construct state sequences, rather than those than construct the parameters (A, B, C) directly from the matrices $(\mathcal{O}_f, \mathcal{K}_p)$. In fact, any closed-loop state-space identification method that estimates state sequences can be directly integrated with the method of Section 4.

Selection of the model dimensions (n, f, p) can either be tuned by hand or with information criteria methods. While dimension selection is outside of the scope of this paper, Bauer (2001), Chiuso (2010) and Larimore (2005) each describe selection of the parameters n, f, and p, respectively. In the case study, we tuned (n, f, p) by hand and validated the chosen state order n with the singular value criterion described by Bauer (2001).

[3] A minimal realization is controllable and observable.

[4] Since the regressors Z_f are correlated with the errors E_f, joint estimation the parameters $(\mathcal{H}_{f,p}, \mathcal{G}_f)$ produces inconsistent estimates. As a result, we have to eliminate the $\mathcal{G}_f Z_f$ term of the model (12) via a "pre-estimation" step. In open-loop subspace methods, a slightly different extended state-space model is used, allowing for consistent estimation of the parameters $(\mathcal{H}_{f,p}, \mathcal{G}_f)$ under open-loop conditions.

[5] N4SID and MOESP methods are weighted least squares problems, whereas CVA is an approximate ML problem.

4. Closed-loop disturbance model identification

In this section, we extend the subspace identification algorithm of Section 3 to identify an augmented disturbance model (3). This is done by first estimating a disturbance sequence that captures the most long-term modeling error, and then re-estimating the noise covariances based on that disturbance sequence. The shaping matrices (B_d, C_d) of the noise model are inconsequential to the algorithm, except that they must obey the offset-free rank condition (7), and that output disturbance models turn out to be computationally advantageous.

4.1. Choosing the disturbance model

As previously discussed, the disturbance model (B_d, C_d) can be chosen to maximize interpretability of the augmented disturbance model (3). We propose general guidelines for choosing the disturbance model below.

- If $\hat{A}$ does not contain integrators, use an output disturbance model.
- If $\hat{A}$ contains integrators and $n_u = n_y$, use an input disturbance model, $(B_d, C_d) = (B, 0)$.
- Otherwise, use some combination of input and output disturbances, i.e. $(B_d, C_d) = (B\tilde{I}_1, \tilde{I}_2)$ where $\tilde{I}_1$ and $\tilde{I}_2$ are diagonal matrices with zeros and ones on the diagonal and collectively n_y nonzero elements.

Models in these forms retain interpretability while ensuring that the offset-free rank condition (7) is satisfied.

4.2. Estimating the disturbance sequence

Given a model of the form (2), a disturbance model (B_d, C_d), and a state sequence $(\tilde{x}(k))$, we treat the disturbance sequence $(d(k))$ as accounting for the long-range model errors. That is, the long-range output is

$$\begin{aligned} y(k) &= \hat{C}\hat{A}^{k-p}\tilde{x}(p) + \sum_{j=p}^{k-1} \hat{C}\hat{A}^{k-j-1}\hat{B}u(j) \\ &\quad + \sum_{j=p}^{k-1} \hat{C}\hat{A}^{k-j-1}(B_d d(j) + w(j)) + C_d d(k) + v(k) \end{aligned}$$

and the *predicted* long-range output is

$$\hat{y}(k) := \hat{C}\hat{A}^{k-p}\tilde{x}(p) + \sum_{j=p}^{k-1} \hat{C}\hat{A}^{k-j-1}\hat{B}u(j) \tag{14}$$

Next, we define the long-range prediction error as $z(k) := y(k) - \hat{y}(k)$, which gives

$$z(k) = \sum_{j=p}^{k-1} \hat{C}\hat{A}^{k-j-1}(B_d d(j) + w(j)) + C_d d(k) + v(k)$$

Rewriting this as a linear model,

$$Z_{N_s}(p) = \mathcal{A} D_{N_s}(p) + \mathcal{B} W_{N_s}(p) + V_{N_s}(p) \tag{15a}$$

$$\mathcal{B} W_{N_s}(p) + V_{N_s}(p) \sim \mathcal{N}(0, \mathcal{V}) \tag{15b}$$

where

$$\mathcal{A} := \begin{bmatrix} C_d & & & \\ \hat{C}B_d & C_d & & \\ \vdots & \ddots & \ddots & \\ \hat{C}\hat{A}^{N-2}B_d & \dots & \hat{C}B_d & C_d \end{bmatrix}$$

$$\mathcal{B} := \begin{bmatrix} 0 & & & \\ B_1 & 0 & & \\ \vdots & \ddots & \ddots & \\ B_{N-1} & \dots & B_1 & 0 \end{bmatrix}$$

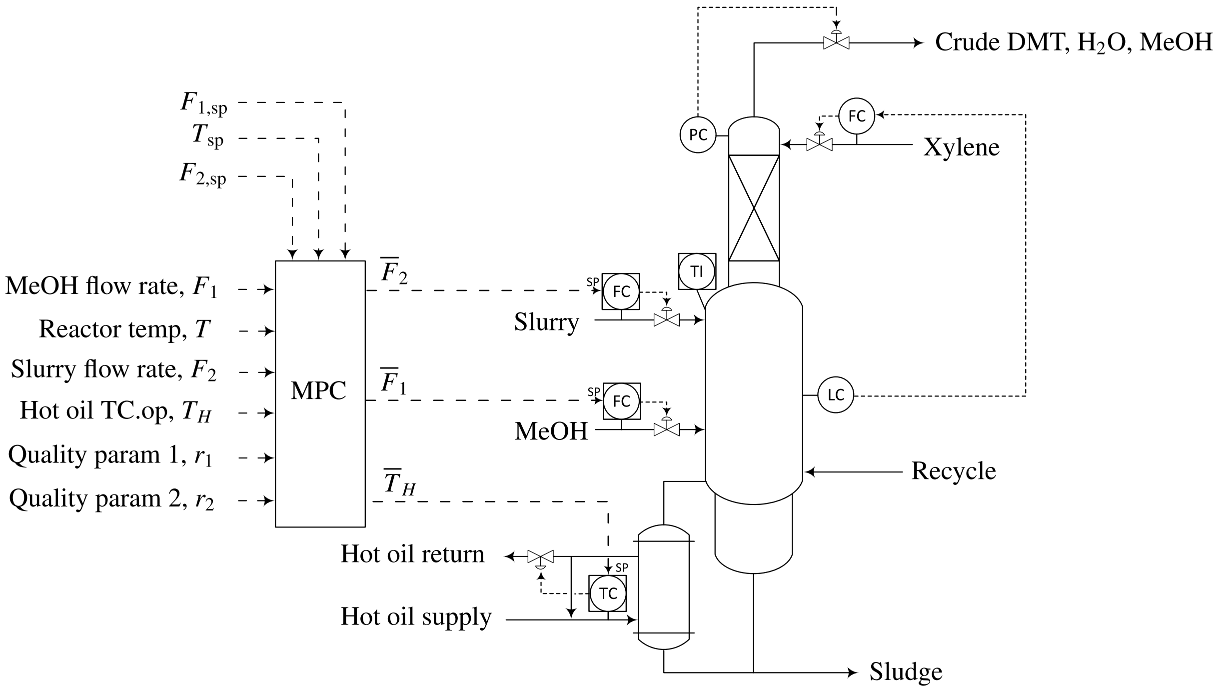

Fig. 2. Schematic of the DMT reactor and MPC control strategy.

$$B_j := \hat{C}\hat{A}^{j-1} \quad \forall j \geq 1$$
$$\mathcal{V} := \mathcal{B}(I \otimes Q_w)\mathcal{B}' + I \otimes R_v$$

The model (15) has the ML estimate (Rao, 1971; Magnus and Neudecker, 2019, p. 313),

$$\hat{D}_{N_s}(p) = (\mathcal{A}'\mathcal{V}_0^{\dagger}\mathcal{A})^{\dagger}\mathcal{A}'\mathcal{V}_0^{\dagger}Z_{N_s}(p) \tag{16}$$

where $\mathcal{V}_0 := \mathcal{V} + \mathcal{A}\mathcal{A}'$. This is an $O(N^3)$ computation with $O(N^2)$ memory requirements. Notice that when $B_d = 0$ and $C_d = I$, we have $\mathcal{A} = I$, $\mathcal{V}_0 = \mathcal{V} + I$ invertible, and

$$(\mathcal{A}'\mathcal{V}_0^{\dagger}\mathcal{A})^{\dagger}\mathcal{A}'\mathcal{V}_0^{\dagger} = \mathcal{V}_0\mathcal{V}_0^{-1} = I$$

Thus the disturbance estimates (16) are equivalently written

$$\hat{d}(k) = z(k) \tag{17}$$

which is an $O(N)$ computation without additional memory requirements. It is clear that whenever the system is free of integrators, the simplified solution (17) is computationally advantageous. A similarity transformation can be used to find the desired disturbance model after the output disturbance model is found (Rajamani et al., 2009).

4.3. Estimating the noise covariances

Given the estimated states and disturbances, one can stack the equations of the model (3) to write a simple covariance estimation problem,

$$\tilde{e}(k) := \begin{bmatrix} \tilde{x}(k+1) \\ \hat{d}(k+1) \\ y(k) \end{bmatrix} - \begin{bmatrix} \hat{A} & B_d & \hat{B} \\ 0 & I & 0 \\ \hat{C} & C_d & 0 \end{bmatrix} \begin{bmatrix} \tilde{x}(k) \\ \hat{d}(k) \\ u(k) \end{bmatrix} \overset{iid}{\sim} \mathcal{N}(0, S_d)$$

The ML estimate of S_d is therefore $\hat{S}_d = S(\tilde{e}, \tilde{e})$ (Anderson, 2003, Thm. 8.2.1). Thus, we have found the complete set of parameters for the augmented disturbance model (3), which concludes our description of the algorithm.

5. Industrial case study

To evaluate the proposed closed-loop identification algorithm and experimental design, a case study was conducted on a reactor at Eastman Chemical's plant in Kingsport, Tennessee. The chosen process is similar to that used in Caveness and Downs (2005). The process produces dimethyl terephthalate (DMT) by reacting terephthalic acid (TPA) with methanol (MeOH). Water is a byproduct of the reaction. The primary equilibrium reaction can be represented as

$$\mathrm{TPA} + 2\mathrm{MeOH} \rightleftharpoons \mathrm{DMT} + 2\mathrm{H_2O}$$

TPA is a solid and enters the reactor in a slurry with methanol, and additional methanol enters as a vapor. The reactor has two phases. The reaction takes place in a liquid phase, and the DMT product, water, excess methanol, and side products leave the reactor as a vapor and move forward to a DMT purification section. Xylene is added as reflux to minimize the carryover of an impurity that results from the half reaction of TPA and methanol. Xylene does not participate in the reaction. A schematic of the reactor is shown in Fig. 2.

The reactor operates under pressure, which is controlled by manipulating a valve in the vapor line. Heat is supplied to the reboiler by circulating hot oil through the shell side of the exchanger. A temperature controller manipulates the flow of hot fluid supplying the circulation loop to control the temperature of the heating fluid entering the reboiler. Liquid level is controlled by manipulating the xylene reflux. Any change in the material balance that affects the composition of methanol in the reactor has a large influence on reactor temperature. Infinite-horizon MPC (4)–(6) is used to control the reactor temperature, T, and the production rate (ultimately set by the slurry feed, F_2) and to maintain the methanol feed, F_1, at a desired rate. The MPC also handles constraints on two quality-control variables, r_1 and r_2, and on the hot oil controller valve position (used to infer a temperature pinch/constraint on hot oil temperature, T_H). The manipulated

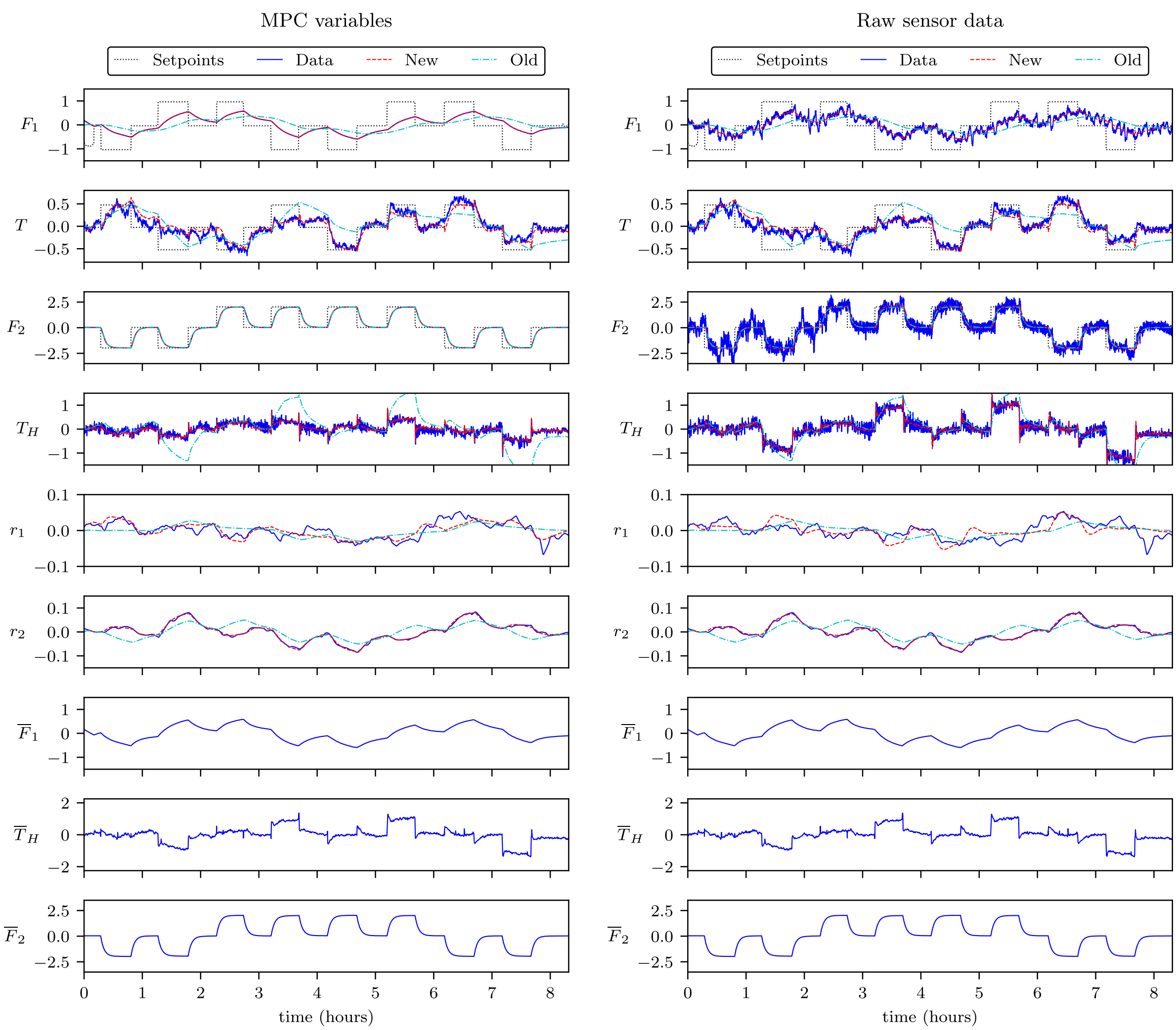

Fig. 3. Controlled variables (top three), other measured variables (middle three), and manipulated variables (bottom three) for the closed-loop identification experiment using (left) the MPC variables ($n = 15$, $f = 5$, $p = 50$) and (right) the raw sensor data ($n = 15$, $f = 5$, $p = 50$). The dotted lines are MPC setpoints, the dot–dashed lines are the predictions of the old MPC model, and the dashed lines are the predictions of the new model. Predictions are long-range projections based on a zero initial state (14).

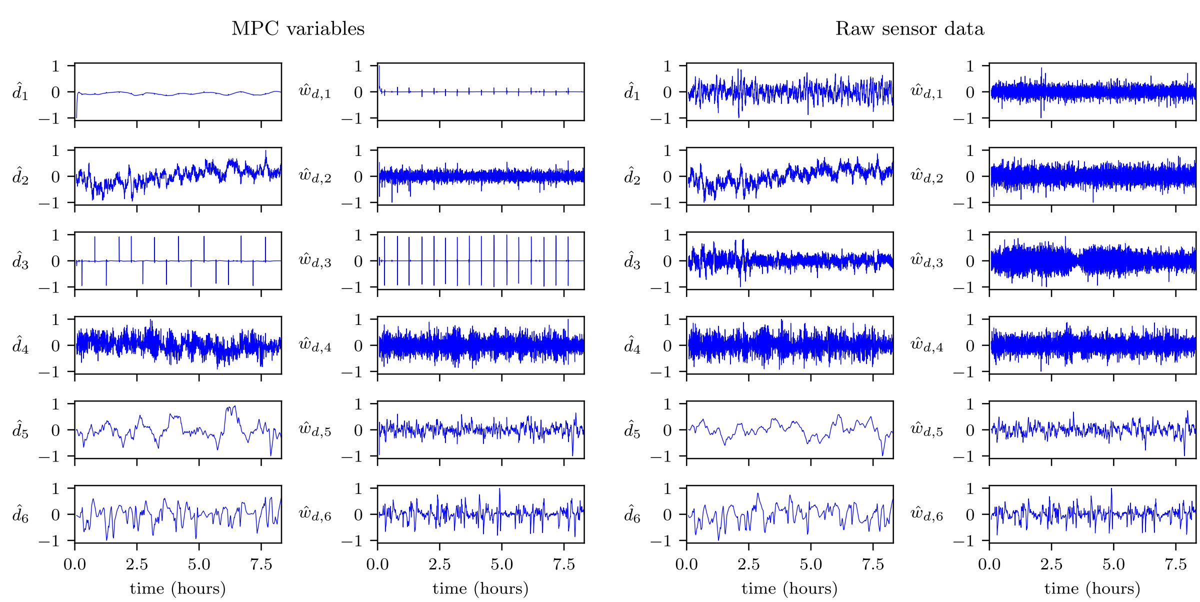

Fig. 4. Disturbance estimates (17) and driving noise estimates $\hat{w}_d = \hat{d}^+ - \hat{d}$ for the unregularized models fit to (left) the MPC variables and (right) raw sensor data. To aid readability, the disturbance estimates were rescaled to have a maximum absolute deviation of 1.

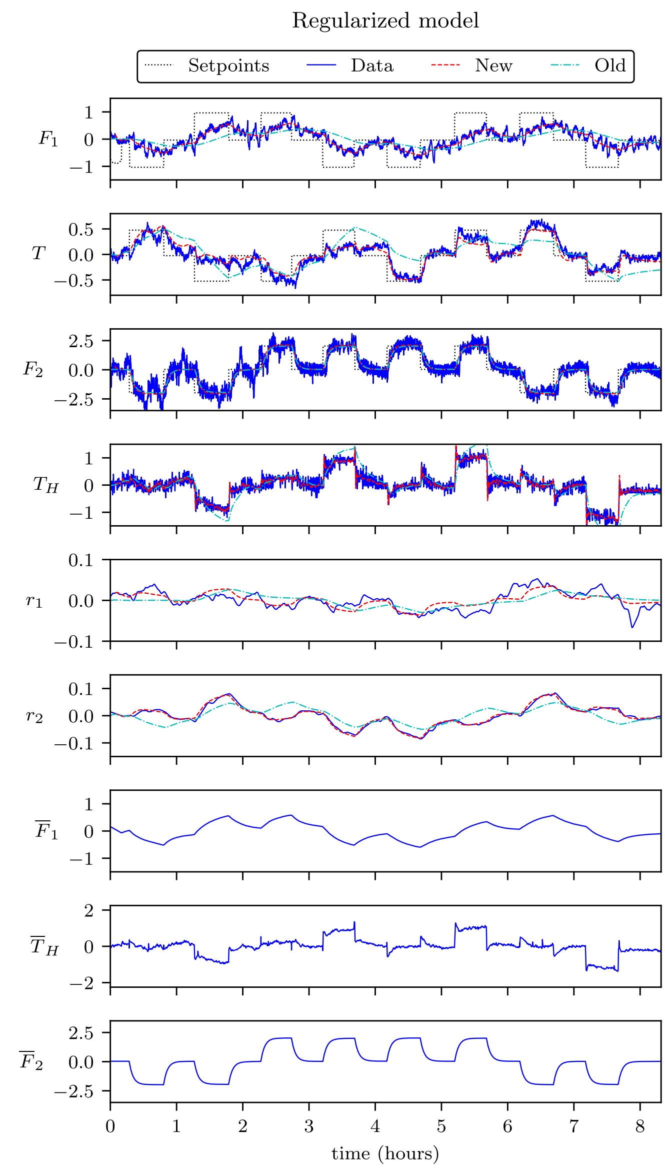

Fig. 5. Controlled variables (top three), other measured variables (middle three), and manipulated variables (bottom three) for the closed-loop identification experiment using the regularized model ($n = 15$, $f = 5$, $p = 50$, $\rho = 10^{-4}\|S(Z_{-p}, Z_{-p})\|^2$, $\mu_1 = 10^{-7}\|S(Z_{-p}, Z_{-p})\|^2$, and $\mu_2 = 10^{-4}\|S(\tilde{x}, \tilde{x})\|^2$). The dotted lines are MPC setpoints, the dot–dashed lines are the predictions of the old MPC model, and the dashed lines are the predictions of the new model. Predictions are long-range projections based on a zero initial state (14).

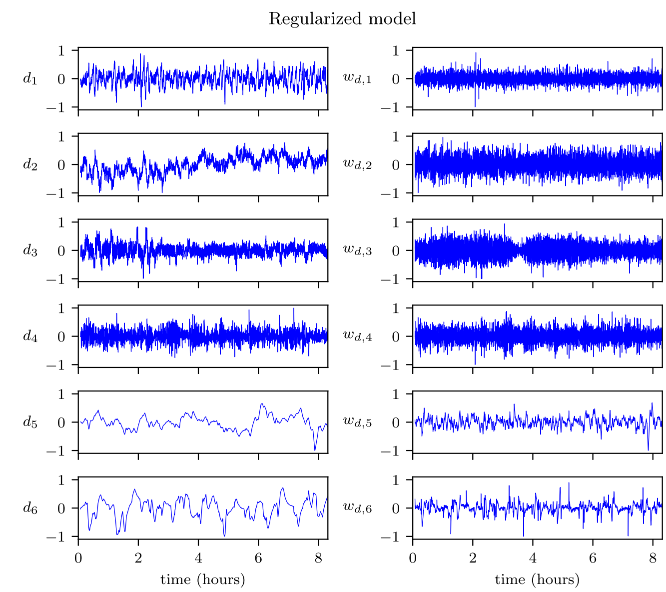

Fig. 6. Disturbance estimates (17) and driving noise estimates $\hat{w}_d = \hat{d}^+ - \hat{d}$ for the regularized model. To aid readability, the disturbance estimates were rescaled to have a maximum absolute deviation of 1.

variables are the PID loop setpoints for the inlet flowrate and utility temperature controllers, denoted $(\overline{F}_1, \overline{F}_2, \overline{T}_H)$.

The control objectives are to achieve offset-free setpoint tracking and disturbance rejection and to avoid violating box constraints on the measured and manipulated variables. For several decades the reactor has run on an MPC designed with a step response model (to be referred to as the "old MPC model") and hand-tuned estimator, as described in Caveness and Downs (2005). The inlet flowrate "measurements" are actually "wrap-around" variables, that is, each flowrate "measurement" is generated by passing the corresponding PID setpoint (the MPC's actuator) through a first-order filter.[6] We refer to these fictitious flowrate "measurements" as the "wrap-around" variables, and the actual flowrate data, collected from the PID layer, as the raw sensor data. We refer to the complete dataset (3 inputs, 6 outputs) formed with the "wrap-around" variables as the "MPC variables" and the complete dataset formed with the sensor data as the "raw sensor data". The MPC runs at a sample time of 5 s.

5.1. Identification

To identify the process, we used a closed-loop experimental design based on pulses to the normal MPC setpoints. Eight setpoint pulses were applied, each lasting about 30 min, with 30 min "rests" between the pulses to allow the process to settle back to the normal operating point. The setpoint pulses correspond to a full factorial design of the three controlled variables.[7] The pulses were designed to keep the manipulated and measured variables within constraints, and they were checked against historical data to ensure production would not be negatively affected. Throughout, models are fit with the algorithm of Sections 3, 4 and Appendix.

"Wrap-around" variables and sensor data. Models were fit to two sets of process data. The first dataset was constructed from the "wrap-around" variables used on the existing MPC, and the corresponding model uses parameters $n = 20$, $f = 5$, and $p = 50$.[8] The second dataset was constructed from the raw sensor data, and the corresponding model uses parameters $n = 15$, $f = 5$, and $p = 50$. In Fig. 3, for each dataset, we plot process data, setpoint changes, and long-range predictions (14)

[6] Given the clarity of hindsight, we would not design the MPC with these fictitious variables. However, our objective in this paper is not to scrutinize the MPC organizational design (that is, the variable choices) but to identify and validate a flexible replacement model via closed-loop experiments. It is worth pointing out that practitioners and academics alike agree that a significant opportunity in MPC performance gains is in improving the organizational structure of implementations (Darby and Nikolaou, 2012).

[7] Because the manipulated and controlled variables form a square system, we could perturb the setpoints without worrying about correlation in the manipulated variables.

[8] Here we violate the assumption, used in Appendix, that $f \geq n$. This condition is only sufficient for producing a rank-n Hankel matrix $\mathcal{H}_{f,p}$. In practice, it is not necessary, so we used the smallest values of (f, p, n) to accurately predict system behavior.

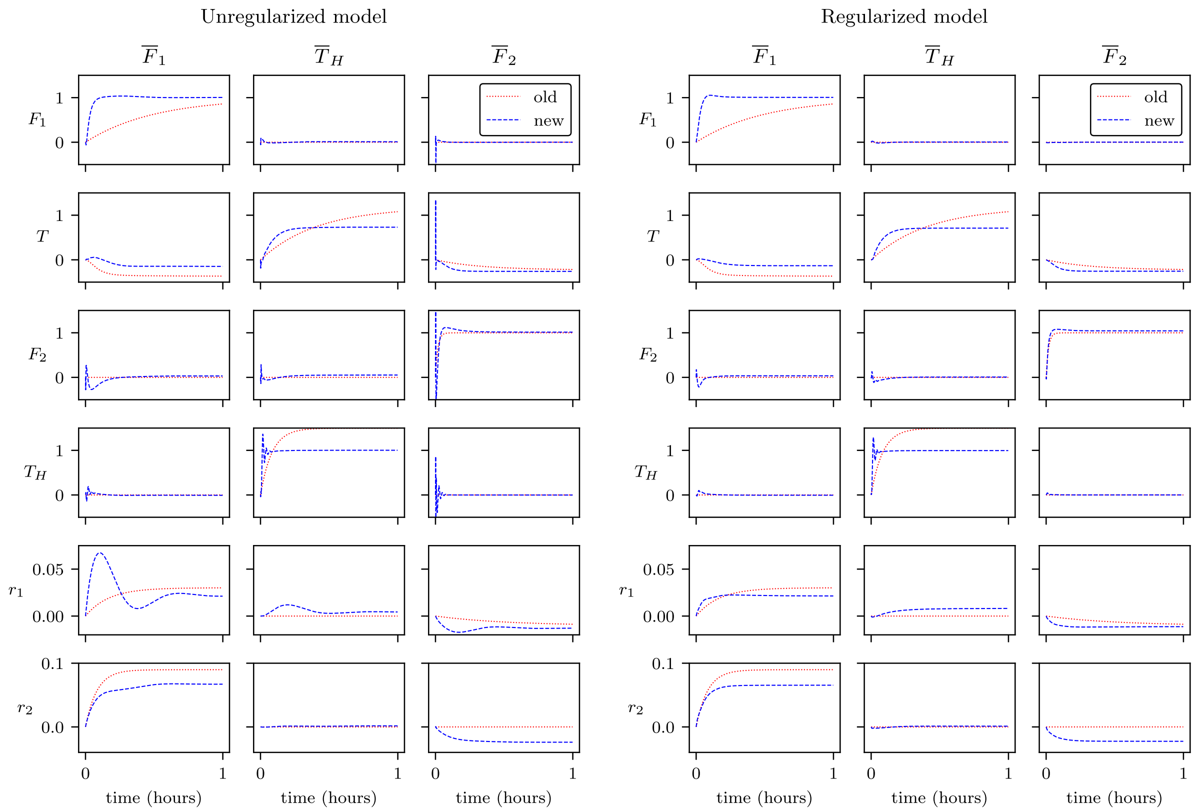

Fig. 7. Step responses of the (left) unregularized and (right) regularized models compared to the step responses of the old MPC model.

of the old and new models. The disturbance estimates (17) and driving noise $\hat{w}_d = \hat{d}^+ - \hat{d}$ for each model is plotted in Fig. 4. From Fig. 4 (left), it is clear that the model fit to the MPC variables is not driven by white noise. This is to be expected; the MPC variables contain outputs that are *not* constructed from sensor data and therefore do not include upstream disturbances affecting, for example, the PID layer dynamics and offset. As the assumptions of the augmented disturbance model (3) are violated, we chose to continue with the model based on raw sensor data, which is clearly driven by white noise (Fig. 4, right). It is worth pointing out that, in the experiment, the temperature failed to reach the second and fourth setpoints. This is due to plant-model mismatch in the old MPC model, as that model incorrectly predicts that the temperature will reach the setpoint. The newly identified models do not make such predictions. Additionally, the first flowrate F_1 never reaches any of the setpoints because it has a low regulator weight relative to that of the temperature. Despite the significant noise present in the raw sensor data, the model fit to this data is no worse at predicting the outputs than the model fit to the MPC variables.

Model regularization. Regularization is a classic technique in statistics and linear algebra used to avoid model over-fitting and ill-conditioning (Tikhonov, 1963; Hoerl and Kennard, 1970b,a). While it is less common in system identification, there is a history of its use for at least three decades (Sjöberg et al., 1993; Johansen, 1997; Chen and Ljung, 2013; Chen et al., 2014). To investigate the possibility of model over-fitting, we also used regularized estimates to produce a model. See Appendix for a derivation of the regularized estimates and the meaning of the regularization parameters. A regularized model was fit to the raw measurement data using parameters $n = 15$, $f = 5$, $p = 50$, $\rho = 10^{-4}\|S(Z_{-p}, Z_{-p})\|^2$, $\mu_1 = 10^{-7}\|S(Z_{-p}, Z_{-p})\|^2$, and $\mu_2 = 10^{-4}\|S(\tilde{x}, \tilde{x})\|^2$. Process data, setpoint changes, and long-range predictions (14), for both old and new models, are plotted in Fig. 5. The disturbance estimates (17) and driving noise $\hat{w}_d = \hat{d}^+ - \hat{d}$ for each model is plotted in Fig. 6.

As a sanity check of the model fits (and to tune the regularization parameters) we plotted the step responses of the unregularized and regularized models (Fig. 7). At a first glance, the long-range predictions in Fig. 3 (right) appear to be representative of the true process dynamics. However, when looking at the step responses of the model (Fig. 7, left) it is clear that there are artifacts and spurious dynamics in the model fit that we speculate is due to over-fitting of the plant model to the disturbance signal in the high frequency range. Regularization takes care of these problems, creating a smoother step response (Fig. 7, right). As such, we chose to update the MPC on the process in Fig. 2 with the regularized model.

5.2. Closed-loop performance

To evaluate the performance of the new MPC model, we used a closed-loop experimental design similar to the one carried out during identification. Again, eight setpoint pulses were applied, each lasting about 30 min, with 30 min "rests" between the pulses to allow the process to settle back to the normal operating point. This time, however, the experiment was carried out over two separate days, switching the MPC model between the two days. Both experiments used the same infinite horizon MPC (4)–(6) with the only difference being the model and estimator gain. It is worth pointing out that, while the new model was fit to the raw sensor data, the MPC uses the "wrap-around" variables in both experiments. As a result, there is a risk the MPC does not respond to disturbances affecting these measurements. The MPC variables and raw sensor data from these experiments are plotted in Figs. 8 and 9, respectively. From these plots, it appears that the F_1 valve needed servicing. However, because feedback was done with the "wrap-around" variables, there was no effect on the closed-loop performance. It is also clear that the old MPC model continues to have difficulties reaching certain temperature setpoints, whereas the new model is confirmed to alleviate these problems. In the new

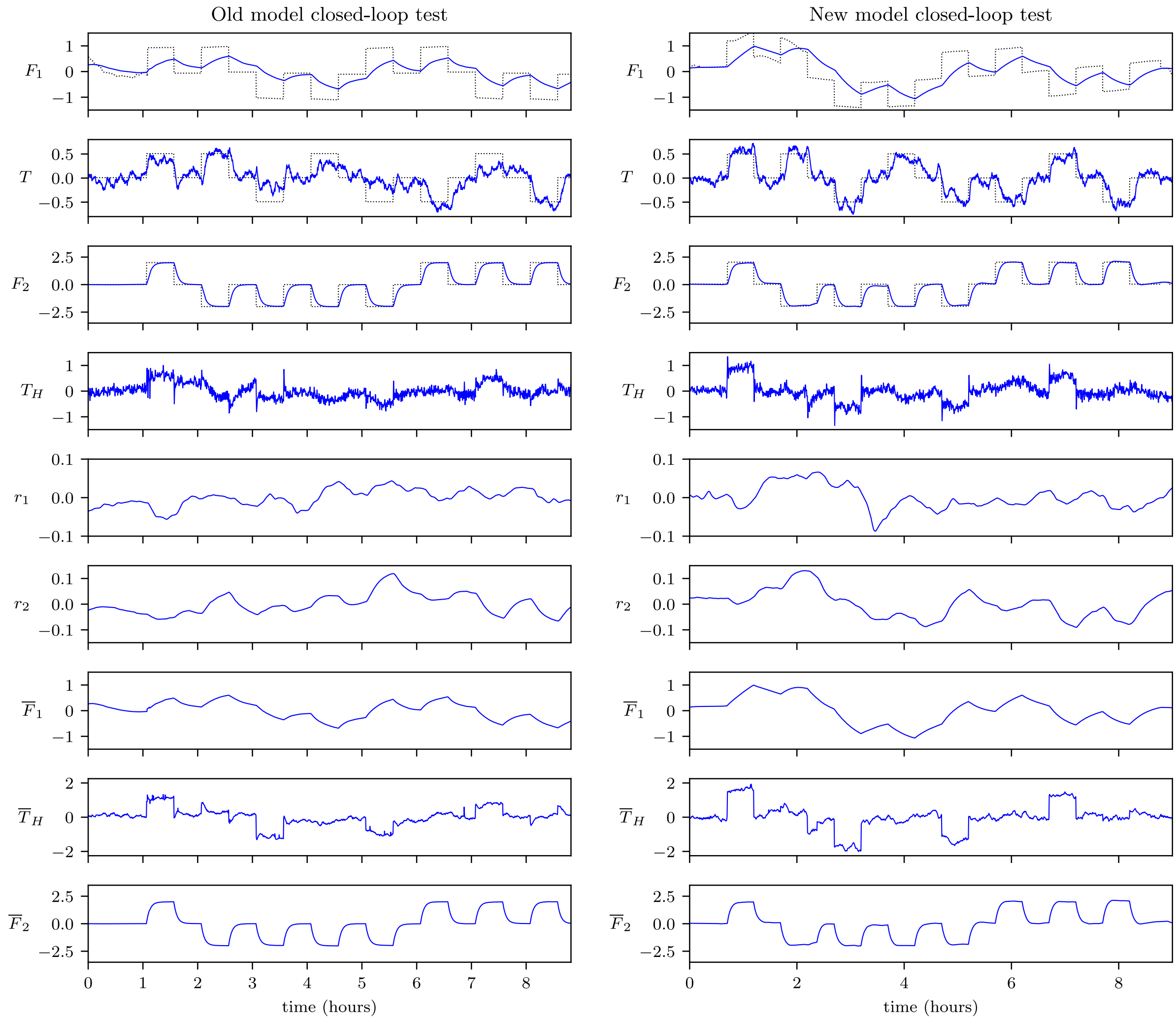

Fig. 8. Closed-loop comparison of the (left) old and (right) new models using the MPC variables.

model, deviations from setpoints are zero mean, so they are likely attributable to process noise and upstream disturbances. Again, both MPC implementations fail to reach F_1 setpoints as this variable has a low regulator weight relative to that of the temperature. The Kalman filtered disturbance estimates (4) for the old and new models (using the MPC variables as feedback) are plotted in Fig. 10. The new model has a much quicker filter gain. This is particularly prevalent in the $\hat{d}_4$ Kalman filter estimate (which corresponds to the T_H measurement), which is slow for the old model but virtually instantaneous for the new model.

To quantify the performance of each MPC, we computed the controlled variable tracking cost,

$$\ell(k) := \|Hy(k) - r_{\mathrm{sp}}(k)\|^2_{Q_y}$$

where $H = \begin{bmatrix} I_3 & 0 \end{bmatrix}$ and $Q_y = \mathrm{diag}(10^{-4}, 1, 10^{-3})$, which is approximately the squared error between T and its setpoint. Tighter control will exhibit a smaller tracking cost $\ell(k)$, on average. It is known that for linear plants and linear controllers without constraints, the tracking cost $\ell(k)$ has a generalized-χ^2 distribution,[9] but if it is time-averaged, it will approach a normal distribution (Zagrobelny et al., 2013). We define the T-lagged average at time k as

$$\langle \ell(k) \rangle_T := \frac{1}{T} \sum_{j=0}^{T-1} \ell(k-j)$$

We compare tracking costs $\ell(k)$ and time-averaged tracking costs $\langle \ell(k) \rangle_{1000}$ and $\langle \ell(k) \rangle_k$ for the old and new models in Fig. 11. It is immediately clear that the new model performed better than the old model; the total average tracking cost (Fig. 11, bottom left) is 38% lower in the new model experiment compared to that of the old model experiment. The cost $\ell(k)$ (Fig. 11, top right) fits the linear control assumptions on generalized-χ^2 distribution. Moreover, the time-averaged cost $\langle \ell(k) \rangle_{1000}$ (Fig. 11, bottom right) is approaching a normal distribution, although there is some residual density near $\ell = 0$ for both experiments. These results suggest the applicability of a statistical performance monitoring scheme such as the one in Zagrobelny et al. (2013).

6. Conclusion

We present and validate a method for identifying linear augmented disturbance models from closed-loop data, which provides all necessary information to design the MPC estimator. The method is based on a nested ML algorithm that first produces a disturbance-free model,

[9] A generalized-χ^2 random variable is generated by taking the quadratic form of a multivariate normal random variable.

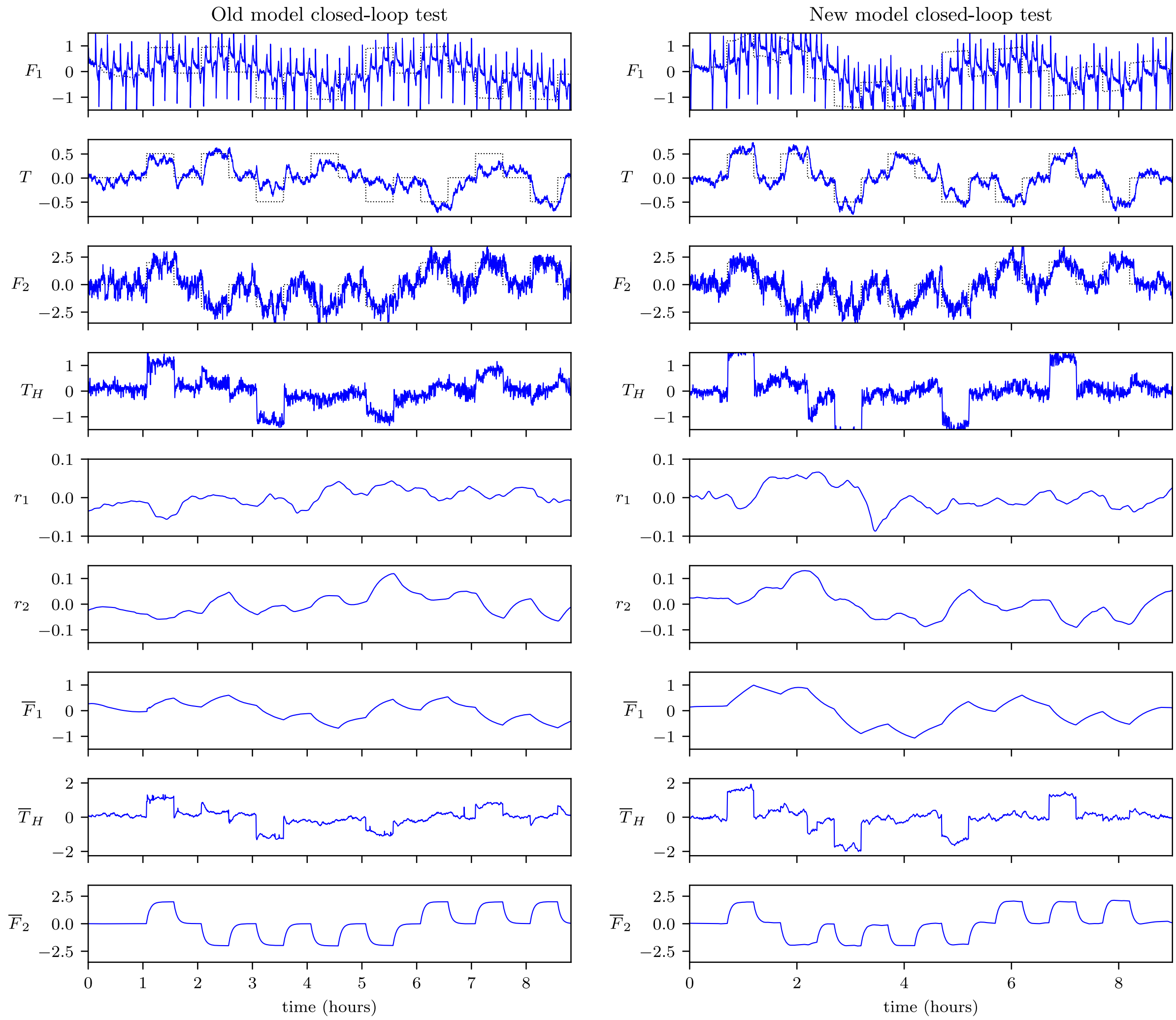

Fig. 9. Closed-loop comparison of the (left) old and (right) new models using the raw process variables as measurements.

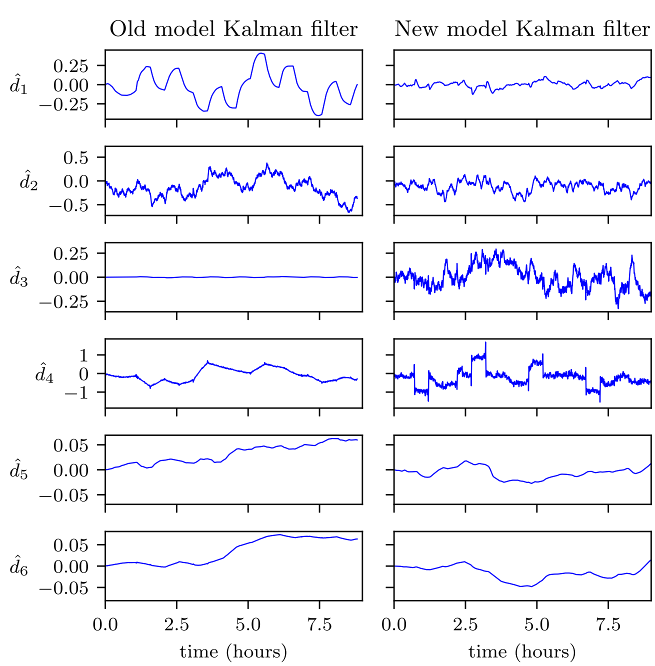

Fig. 10. Kalman filtered disturbances from the closed-loop tests of the (left) old and (right) new models using the MPC variables as measurements.

then estimates disturbance sequences, and finally estimates noise covariances. The method is tested on an existing reactor at Eastman Chemical's Kingsport, Tennessee plant. We show the method is able to fit process models from fictitious measurements that may be present on some legacy MPC implementations. The importance of regularization to avoid over-fitting of high-frequency disturbances is investigated. The ability to use closed-loop data allows practitioners to safely and cheaply identify and re-identify their processes. The model is validated in a closed-loop test and the tracking error recorded for the duration of the experiment. The models produce generalized χ^2-distributed tracking errors that validate the use of statistical performance monitoring algorithms such as that of Zagrobelny et al. (2013). Most importantly, the new model outperforms the old model in both qualitative metrics (reaching setpoints, speed of the estimator) and quantitative metrics (38% reduction in tracking error).

There are many possibilities in future case studies of this technology, including comparisons across competing methods (ALS, EM) and computational studies. The use of a tracking metric to validate the new model performance is limited in scope to tracking MPC, but this work can be extended using the work of Zanon et al. (2016, 2017), providing a practical way to implement economic MPCs in a linear–quadratic framework. While the validity of linearized economic MPC, with exact models, is theoretically supported, there are no case studies or theoretical analysis on employing these linear economic MPCs with identified models. Other considerations could be taken into account during the identification of the standard model (2), including measured

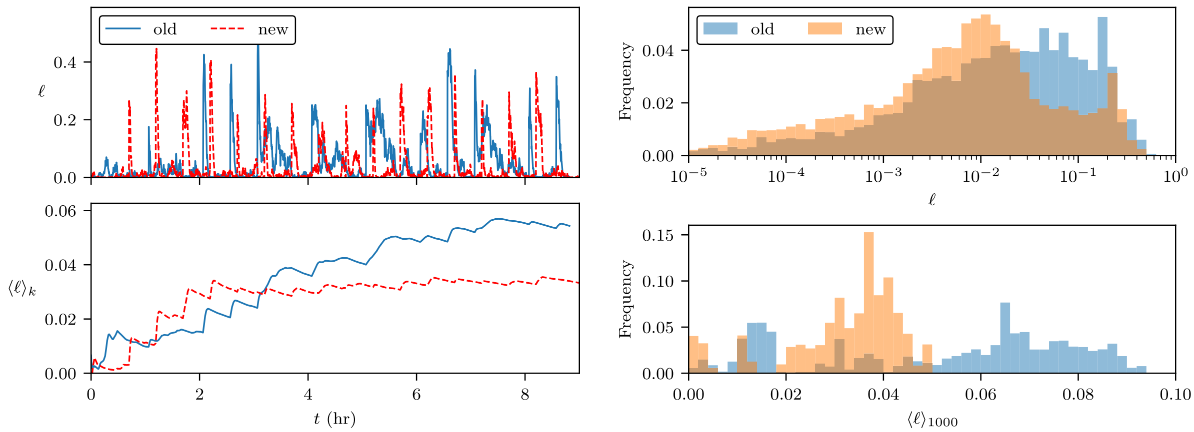

Fig. 11. Closed-loop performance model of the old and new models using the MPC variables as measurements. (Top left) Output tracking error $\ell(k)$, (bottom left) running average of the tracking error $\langle \ell(k) \rangle_k$, (top right) histogram of the output tracking error $\ell(k)$, (bottom right) histogram of the output tracking error moving average $\langle \ell(k) \rangle_{1000}$.

disturbances and flowsheet structure. With a complete set of parameter estimates and uncertainties, more advanced offset-free MPC monitoring schemes can be envisioned. Finally, future research may be directed towards developing a theoretical understanding of the disturbance model identification method and how it relates to competing methods (ALS, EM).

CRediT authorship contribution statement

Steven J. Kuntz: Conceptualization, Methodology, Software, Investigation, Data curation, Writing – original draft, Visualization. **James J. Downs:** Conceptualization, Investigation, Writing – review & editing. **Stephen M. Miller:** Conceptualization, Writing – review & editing, Supervision. **James B. Rawlings:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The raw data required to reproduce the above findings is proprietary and cannot be shared at this time.

Acknowledgment

This work was supported by the National Science Foundation (NSF) under Grant 2138985 and the industrial members of the Texas Wisconsin California Control Consortium (TWCCC).

Appendix. Closed-loop subspace identification

In this appendix, we continue the explanation of the closed-loop subspace method that began in Section 3. First, we derive the (regularized) ML estimator for the HOARX model (13). Second, we derive a ML estimator for the extended state-space model (12). Last, we derive the (regularized) ML estimator for the model (2), corresponding to the joint ML problem (10).

A.1. Estimating the Markov parameters

The likelihood function corresponding to the ARX model (13) is an approximation of the likelihood function $L_N(\theta)$,

$$\begin{aligned} L_N(\theta) &\approx L_N^{\mathrm{ARX}}(C\mathcal{K}_{\bar{p}}, R_e) \\ &:= \sum_{k=\bar{p}}^{N-1} \ln p(y(k)|Z_{-\bar{p}}(k), C\mathcal{K}_{\bar{p}}, R_e) \\ &\propto -\frac{N-\bar{p}}{2} \ln |R_e| - \frac{1}{2} \sum_{k=\bar{p}}^{N-1} \|y(k) - C\mathcal{K}_{\bar{p}} Z_{-\bar{p}}(k)\|^2_{R_e^{-1}} \end{aligned}$$

The approximate ML problem

$$\max_{C\mathcal{K}_{\bar{p}}, R_e > 0} L_N^{\mathrm{ARX}}(C\mathcal{K}_{\bar{p}}, R_e)$$

is a standard multivariate regression problem with closed-form solution,

$$\widehat{C\mathcal{K}_{\bar{p}}} = S(y, Z_{-\bar{p}}) S^{-1}(Z_{-\bar{p}}, Z_{-\bar{p}})$$

We omit the solution for $\hat{R}_e$ because it is inconsequential to the rest of the algorithm. The ARX model is an overparameterization of the model (2), so it is beneficial to regularize the coefficients, trading a biased estimate for reduced variance,[10]

$$\max_{C\mathcal{K}_{\bar{p}}, R_e > 0} L_N^{\mathrm{ARX}}(C\mathcal{K}_{\bar{p}}, R_e) - \frac{\rho}{2} \operatorname{tr}(R_e^{-1} C\mathcal{K}_{\bar{p}} (C\mathcal{K}_{\bar{p}})') \tag{A.1}$$

which results in the regularized estimates,

$$\widehat{C\mathcal{K}_{\bar{p}}} = S(y, Z_{-\bar{p}}) \left[S(Z_{-\bar{p}}, Z_{-\bar{p}}) + (\rho/N_s) I \right]^{-1} \tag{A.2}$$

The estimates (A.2) are unbiased when $\rho = 0$ and consistent for all $\rho \geq 0$.[11] Moreover, the estimate errors $\mathcal{E}_{\mathrm{ARX}} := \widehat{C\mathcal{K}_{\bar{p}}} - C\mathcal{K}_{\bar{p}} = [\hat{G}_1 - G_1, \ldots, \hat{G}_{\bar{p}} - G_{\bar{p}}]$ are independent of the innovation sequence $e(k)$ and regression vectors $Z_{-\bar{p}}(k)$.

A.2. Estimating the state sequence

It turns out that the likelihood of the extended state-space model (12), even though the errors $E_f(k)$ are serially correlated, is an approximation of the likelihood in the ML problem (9) (Larimore, 1997).

[10] The regularizer here is close to using the prior $(C\mathcal{K}_{\bar{p}})_i \overset{iid}{\sim} \mathcal{N}(0, \rho^{-1} R_e)$ where $(C\mathcal{K}_{\bar{p}})_i$ denotes the ith column of $C\mathcal{K}_{\bar{p}}$, but to be equivalent, we would also need to add $(-\rho(n_u + n_y)\bar{p}/2) \ln |R_e|$ to the likelihood.

[11] This neglects numerical errors introduced by the approximation $A_K^p \approx 0$.

Assume that $N \gg f, p$. Then, for each $s \in \mathbb{I}_{p:p+f-1}$, we can use successive conditioning to write the likelihood as

$$L_N(\theta) \approx \ln p(Y_{(M_s-1)f+s}(s)|Y_{-s}(s), U_{(M_s-1)f+s}(s), \theta) \approx \sum_{m=0}^{M_s-1} \ell_{f,p}(mf+s)$$

where $M_s := \lfloor (N-s)/f \rfloor$ and

$$\ell_{f,p}(k) := \ln p(Y_f(k)|Z_{-p}(k), U_f(k), \mathcal{H}_{f,p}, \mathcal{G}_f, \mathcal{R}_f)$$

Terms at times $k \in \mathbb{I}_{0:N-1} \setminus \mathbb{I}_{s:M_s f+s-1}$ can be dropped because of the assumption that $N \gg f, p$. Taking the average over s gives

$$L_N(\theta) \approx \frac{1}{f} \sum_{s=p}^{p+f-1} \sum_{m=0}^{M_s-1} \ell_{f,p}(mf+s) = \frac{1}{f} \sum_{k=p}^{N-f} \ell_{f,p}(k) =: \frac{1}{f} L_N^{\text{ESS}}(\mathcal{H}_{f,p}, \mathcal{G}_f, \mathcal{R}_f)$$

For closed-loop data, the signals Z_f and E_f are correlated, which may introduce bias into the estimates if all the parameters $(\mathcal{H}_{f,p}, \mathcal{G}_f, \mathcal{R}_f)$ are estimated simultaneously (Qin, 2006). Noting that the future data coefficients $\mathcal{G}_f$ is simply a linear function of the ARX coefficients, i.e. $\mathcal{G}_f = \mathcal{L}(C\mathcal{K}_{\overline{p}})$, the future data term in the model (12) can be eliminated as follows,

$$\tilde{Y}_f(k) := Y_f(k) - \hat{\mathcal{G}}_f Z_f(k) \approx \mathcal{H}_{f,p} Z_{-p}(k) + \mathcal{E}_{\text{ESS}}(k) \tag{A.3}$$

where $\hat{\mathcal{G}}_f := \mathcal{L}(\widehat{C\mathcal{K}_{\overline{p}}})$, and $\mathcal{E}_{\text{ESS}} := \mathcal{L}(\mathcal{E}_{\text{ARX}})Z_f + E_f$ is zero-mean since $\mathcal{E}_{\text{ARX}}$ and Z_f are independent. Importantly, the signals Z_{-p} and $\mathcal{E}_{\text{ESS}}$ are uncorrelated, so the parameters $(\mathcal{H}_{f,p}, \mathcal{R}_f)$ can be estimated without bias. The corresponding likelihood function is

$$L_N^{\text{ESS}}(\mathcal{H}_{f,p}, \hat{\mathcal{G}}_f, \mathcal{R}_f) \propto -\frac{N-f-p+1}{2} \ln |\mathcal{R}_f| - \frac{1}{2} \sum_{k=p}^{N-f} \|\tilde{Y}_f(k) - \mathcal{H}_{f,p} Z_{-p}(k)\|^2_{\mathcal{R}_f^{-1}}$$

and the resulting ML problem,

$$\max_{\operatorname{rank} \mathcal{H}_{f,p}=n, \mathcal{R}_f>0} L_N^{\text{ESS}}(\mathcal{H}_{f,p}, \hat{\mathcal{G}}_f, \mathcal{R}_f) \tag{A.4}$$

corresponds to a rank-reduced regression.[12] According to Larimore (1983) and Anderson (1999), the ML problem (A.4) has a closed-form solution,

$$\hat{\mathcal{H}}_{f,p} = S(\tilde{Y}_f, Z_{-p}) J_n' J_n$$

where J_n denotes the first n rows of $J = U' S^{-1/2}(Z_{-p}, Z_{-p})$, and U are the left singular vectors of the following singular value decomposition,

$$S^{-1/2}(Z_{-p}, Z_{-p}) S(Z_{-p}, Y_f) S^{-1/2}(Y_f, Y_f) = USV'$$

Given these estimates, we have the rank factorization $\hat{\mathcal{H}}_{f,p} = \hat{\mathcal{O}}_f \hat{\mathcal{K}}_p$ where $\hat{\mathcal{O}}_f = S(\tilde{Y}_f, Z_{-p}) J_n'$ and $\hat{\mathcal{K}}_p = J_n$. Moreover, the estimate $\hat{\mathcal{K}}_p$ is a consistent and asymptotically normal estimator of $\mathcal{K}_p$ (up to similarity transformation) (Anderson, 1999). Therefore, we have consistent and asymptotically normal estimates of the states,

$$\tilde{x} = J_n Z_{-p} \tag{A.5}$$

A.3. Estimating the state-space parameters

Finally, we estimate the state-space parameters θ by solving the ML problem (10). We have the following likelihood function,

$$L_N^{\text{SS}}(\tilde{X}_{N_s+1}(p), \theta) \propto -\frac{N_s}{2}(\ln |Q_w| + \ln |R_v|) - \frac{1}{2} \sum_{k=p}^{N-1} \left[\|\tilde{x}(k+1) - A\tilde{x}(k) - Bu(k)\|^2_{Q_w^{-1}} + \|y(k) - C\tilde{x}(k)\|^2_{R_v^{-1}} \right]$$

In practice, we have found that the state estimates (A.5) may contain spurious, unwanted dynamics, so we may regularize this objective in a similar manner to the ARX problem (A.1),

$$\max_{\theta} L_N^{\text{SS}}(\tilde{X}_{N_s+1}(p), \theta) - \frac{\mu_1}{2} \operatorname{tr}(Q_w^{-1}(AA' + BB')) - \frac{\mu_2}{2} \operatorname{tr}(R_v^{-1} CC') \tag{A.6}$$

where $\mu_1, \mu_2 > 0$. According to Anderson (2003, Thm. 8.2.1), the regularized estimates are

$$\begin{bmatrix} \hat{A} & \hat{B} \end{bmatrix} = S(\tilde{x}^+, t) \left[S(t,t) + (\mu_1/N_s) I \right]^{-1} \tag{A.7a}$$

$$\hat{C} = S(y, \tilde{x}) \left[S(\tilde{x}, \tilde{x}) + (\mu_2/N_s) I \right]^{-1} \tag{A.7b}$$

$$\hat{Q}_w = S(\tilde{x}^+, \tilde{x}^+) - S(\tilde{x}^+, t) \left[S(t,t) + (\mu_1/N_s) I \right]^{-1} S(t, \tilde{x}^+) \tag{A.7c}$$

$$\hat{R}_v = S(y,y) - S(y,\tilde{x}) \left[S(\tilde{x},\tilde{x}) + (\mu_2/N_s) I \right]^{-1} S(\tilde{x}, y) \tag{A.7d}$$

Since $\tilde{x}$ are consistent estimates and independent of the errors (w, v), the estimates (A.7) are consistent. This completes the closed-loop identification of the model (2) from an input–output sequence.

References

Akaike, H., 1974. A new look at statistical model identification. IEEE Trans. Automat. Control 19 (6), 716–723.

Anderson, T.W., 1999. Asymptotic distribution of the reduced rank regression estimator under general conditions. Ann. Statist. 27 (4), 1141–1154. http://dx.doi.org/10.1214/aos/1017938918.

Anderson, T.W., 2003. An Introduction to Multivariate Statistical Analysis, third ed. John Wiley & Sons, New York.

Arnold, T.J., Rawlings, J.B., 2022. Tractable calculation and estimation of the optimal weighting for autocovariance least squares. IEEE Trans. Automat. Control 67 (11), 6045–6060. http://dx.doi.org/10.1109/TAC.2021.3124193.

Bauer, D., 2001. Order estimation for subspace methods. Automatica 37 (10), 1561–1573.

Canney, W.M., 2003. The future of advanced process control promises more benefits and sustained value. Oil Gas J. 101 (16), 48–54.

Caveness, M.H., Downs, J.J., 2005. Reactor control using infinite horizon model predictive control. In: AIChE Spring Meeting. Atlanta, GA.

Chen, T., Andersen, M.S., Ljung, L., Chiuso, A., Pillonetto, G., 2014. System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques. IEEE Trans. Automat. Control 59 (11), http://dx.doi.org/10.1109/TAC.2014.2351851.

Chen, T., Ljung, L., 2013. Implementation of algorithms for tuning parameters in regularized least squares problems in system identification. Automatica 49 (7), 2213–2220. http://dx.doi.org/10.1016/j.automatica.2013.03.030.

Chiuso, A., 2010. On the asymptotic properties of closed-loop CCA-type subspace algorithms: Equivalence results and role of the future horizon. IEEE Trans. Automat. Control 55 (3), 634–649.

Clarke, D.W., 1991. Adaptive generalized predictive control. In: Proceedings of the Fourth International Conference on Chemical Process Control. pp. 395–417.

Clarke, D.W., Mohtadi, C., Tuffs, P.S., 1987a. Generalized predictive control—Part I. The basic algorithm. Automatica 23 (2), 137–148.

Clarke, D.W., Mohtadi, C., Tuffs, P.S., 1987b. Generalized predictive control—Part II. Extensions and interpretations. Automatica 23 (2), 149–160.

Cutler, C.R., Ramaker, B.L., 1980. Dynamic matrix control—A computer control algorithm. In: Proceedings of the Joint Automatic Control Conference.

Darby, M.L., Nikolaou, M., 2012. MPC: Current practice and challenges. Control Eng. Pract. 20 (4), 328–342.

Gong, G., Samaniego, F.J., 1981. Pseudo maximum likelihood estimation: Theory and applications. Ann. Statist. 9 (4), 861–869.

Harrison, C.A., Qin, S.J., 2009. Discriminating between disturbance and process model mismatch in model predictive control. J. Process Control 19 (10), 1610–1616.

Hespanha, J.P., 2018. Linear Systems Theory, second ed. Princeton University Press, Princeton and Oxford.

Ho, B.L., Kalman, R.E., 1966. Efficient construction of linear state variable models from input/output functions. Regelungstechnik 14, 545–548.

Hoerl, A.E., Kennard, R.W., 1970a. Ridge regression: Applications to nonorthogonal problems. Technometrics 12 (1), 69–82. http://dx.doi.org/10.1080/00401706.1970.10488635.

Hoerl, A.E., Kennard, R.W., 1970b. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67. http://dx.doi.org/10.1080/00401706.1970.10488634.

Jansson, M., 2003. Subspace identification and ARX modeling. IFAC Proc. Vol. 36 (16), 1585–1590.

[12] The rank constraint is a consequence of $(\mathcal{O}_f, \mathcal{C}_p)$ showing up in the regression model as the product $\mathcal{H}_{f,p} := \mathcal{O}_f \mathcal{C}_p$, where we assume $(\mathcal{O}_f, \mathcal{C}_p)$ are both rank-n so the states come from a minimal realization.

Johansen, T.A., 1997. On Tikhonov regularization, bias and variance in nonlinear system identification. Automatica 33 (3), 441–446. http://dx.doi.org/10.1016/S0005-1098(96)00168-9.

Kheradmandi, M., Mhaskar, P., 2018. Model predictive control with closed-loop re-identification. Comput. Chem. Eng. 109, 249–260.

Kuntz, S.J., Downs, J.J., Miller, S.M., Rawlings, J.B., 2023. An industrial case study on the combined identification and offset-free model predictive control of a chemical process. In: FOCAPO/CPC 2023, January 8-12, 2023. San Antonio, Texas.

Kuntz, S.J., Rawlings, J.B., 2022. Maximum likelihood estimation of linear disturbance models for offset-free model predictive control. In: American Control Conference. Atlanta, GA, pp. 3961–3966.

Kwakernaak, H., Sivan, R., 1972. Linear Optimal Control Systems. John Wiley and Sons, New York.

Larimore, W.E., 1983. System identification, reduced-order filtering and modeling via canonical variate analysis. In: Proceedings of the American Control Conference. San Fransisco, CA, pp. 445–451.

Larimore, W.E., 1997. Optimal reduced rank modeling, prediction, monitoring, and control using canonical variate analysis. In: Proceedings of the IFAC 1997 International Symposium on Advanced Control. Banff, Canada, pp. 61–66.

Larimore, W.E., 2005. Maximum likelihood subspace identification for linear, nonlinear, and closed-loop systems. In: Proceedings of the 2005 American Control Conference. IEEE, pp. 2305–2319.

Lee, J.H., Morari, M., García, C.E., 1994. State-space interpretation of model predictive control. Automatica 30 (4), 707–717.

Lee, J.H., Yu, Z.H., 1994. Tuning of model predictive controllers for robust performance. Comput. Chem. Eng. 18 (1), 15–37.

Ljung, L., 1999. System Identification: Theory for the User, second ed. Prentice Hall, New Jersey.

Lundström, P., Lee, J.H., Morari, M., Skogestad, S., 1995. Limitations of dynamic matrix control. Comput. Chem. Eng. 19, 409–421.

Magnus, J.R., Neudecker, H., 2019. Matrix Differential Calculus with Applications in Statistics and Econometrics, third ed. John Wiley & Sons Ltd., Hoboken, NJ.

Marquis, P., Broustail, J.P., 1988. SMOC, a bridge between state space and model predictive controllers: Application to the automation of a hydrotreating unit. In: McAvoy, T.J., Arkun, Y., Zafiriou, E. (Eds.), Proceedings of the 1988 IFAC Workshop on Model Based Process Control. Pergamon Press, Oxford, pp. 37–43.

Morari, M., Maeder, U., 2012. Nonlinear offset-free model predictive control. Automatica 48 (9), 2059–2067.

Muske, K.R., Badgwell, T.A., 2002. Disturbance modeling for offset-free linear model predictive control. J. Process Control 12 (5), 617–632.

Odelson, B.J., Rajamani, M.R., Rawlings, J.B., 2006. A new autocovariance least-squares method for estimating noise covariances. Automatica 42 (2), 303–308.

Pannocchia, G., De Luca, A., 2012. Performance degradation diagnosis and remedies in offset-free MPC. In: 2012 American Control Conference (ACC). pp. 429–434. http://dx.doi.org/10.1109/ACC.2012.6314957.

Pannocchia, G., Rawlings, J.B., 2003. Disturbance models for offset-free MPC control. AIChE J. 49 (2), 426–437.

Qin, S.J., 2006. An overview of subspace identification. Comput. Chem. Eng. 30, 1502–1513.

Qin, S.J., Badgwell, T.A., 2003. A survey of industrial model predictive control technology. Control Eng. Pract. 11 (7), 733–764.

Rajamani, M.R., Rawlings, J.B., 2009. Estimation of the disturbance structure from data using semidefinite programming and optimal weighting. Automatica 45 (1), 142–148.

Rajamani, M.R., Rawlings, J.B., Qin, S.J., 2009. Achieving state estimation equivalence for misassigned disturbances in offset-free model predictive control. AIChE J. 55 (2), 396–407.

Rao, C.R., 1971. Unified theory of linear estimation. Sankhyā A 33 (4), 371–394.

Rao, C.R., 1973. Linear Statistical Inference and Its Applications, second ed. John Wiley and Sons, New York.

Rawlings, J.B., Mayne, D.Q., Diehl, M.M., 2020. Model Predictive Control: Theory, Design, and Computation, 2nd, Paperback ed. Nob Hill Publishing, Santa Barbara, CA, ISBN: 978-0-9759377-5-4, p. 770.

Richalet, J., Rault, A., Testud, J.L., Papon, J., 1978. Model predictive heuristic control: Applications to industrial processes. Automatica 14, 413–428.

Sjöberg, J., McKelvey, T., Ljung, L., 1993. On the use of regularization in system identification. IFAC Proc. Vol. 26 (2), 75–80. http://dx.doi.org/10.1016/S1474-6670(17)48226-7.

Sun, Z., Zhao, Y., Qin, S.J., 2011. Improving industrial MPC performance with data-driven disturbance modeling. In: 2011 50th IEEE Conference on Decision and Control and European Control Conference. pp. 1922–1927. http://dx.doi.org/10.1109/CDC.2011.6161469.

Tikhonov, A.N., 1963. On the solution of ill-posed problems and the method of regularization. In: Doklady Akademii Nauk. Russian Academy of Sciences, pp. 501–504.

Van Overschee, P., De Moor, B., 1994. N4SID: subspace algorithms for the identification of combined deterministic-stochastic systems. Automatica 30 (1), 75–93.

Van Overschee, P., De Moor, B., 1995. A unifying theorem for three subspace system identification algorithms. Automatica 31 (12), 1853–1864.

Verhaegen, M., 1994. Identification of the deterministic part of MIMO state space models given in innovations form from input-output data. Automatica 30 (1), 61–74.

Yousfi, C., Tournier, R., 1991. Steady-state optimization inside model predictive control. In: Proceedings of American Control Conference. p. 1866.

Zagrobelny, M.A., Ji, L., Rawlings, J.B., 2013. Quis custodiet ipsos custodes? Annu. Rev. Control 37, 260–270.

Zagrobelny, M.A., Rawlings, J.B., 2015. Identifying the uncertainty structure using maximum likelihood estimation. In: American Control Conference. Chicago, IL, pp. 422–427.

Zanon, M., Gros, S., Diehl, M., 2016. A tracking MPC formulation that is locally equivalent to economic MPC. J. Process Control 45, 30–42.

Zanon, M., Gros, S., Diehl, M., 2017. A periodic tracking MPC that is locally equivalent to periodic economic MPC. IFAC–PapersOnLine 50 (1), 10711–10716.

Zhu, Y., 2006. System identification for process control: Recent experience and outlook. IFAC Proc. Vol. 39 (1), 20–32. http://dx.doi.org/10.3182/20060329-3-AU-2901.00003.