

Constrained Process Monitoring: Moving-Horizon Approach

Christopher V. Rao

Dept. of Bioengineering, University of California, Berkeley, CA 94720.

James B. Rawlings

Dept. of Chemical Engineering, University of Wisconsin-Madison, Madison, WI 53706-1691

Moving-horizon estimation (MHE) is an optimization-based strategy for process monitoring and state estimation. One may view MHE as an extension for Kalman filtering for constrained and nonlinear processes. MHE, therefore, subsumes both Kalman and extended Kalman filtering. In addition, MHE allows one to include constraints in the estimation problem. One can significantly improve the quality of state estimates for certain problems by incorporating prior knowledge in the form of inequality constraints. Inequality constraints provide a flexible tool for complementing process knowledge. One also may use inequality constraints as a strategy for model simplification. The ability to include constraints and nonlinear dynamics is what distinguishes MHE from other estimation strategies. Both the practical and theoretical issues related to MHE are discussed. Using a series of example monitoring problems, the practical advantages of MHE are illustrated by demonstrating how the addition of constraints can improve and simplify the process monitoring problem.

Introduction

This article discusses the dynamic-inference problem using a state representation, also referred to as a dynamic-state estimation problem. Many control and monitoring systems are based on state-space models. The state is a natural construct when modeling chemical and biological processes, because it compactly summarizes the past information needed to understand the future behavior of the process. For example, temperature, pressure, and concentrations make up the state of a single-phase chemically reactive system. Whether full spatial or simple functional representations such as lumping are employed depends on the accuracy required. However, rarely is the state directly available from the process measurements, and the state typically needs to be inferred from secondary process measurements or a measurable subset of the state. For example, the average molecular weight of many polymer systems is inferred from viscosity measurements. Also, the concentration in a simple chemically reactive system may be inferred from the reactor temperature, a more easily measured state variable.

For a subset of problems, one possesses insights in addition to physical laws and empirical correlations in the form of inequality constraints on the process uncertainties and state variables. For example, many process uncertainties, such as model parameters and process disturbances, are bounded. State variables, such as temperature and concentration, are almost always positive and bounded. These constraints, unlike the process uncertainties, are implicitly enforced by the physical model of the process. However, when one considers approximate models, this implicit enforcement may break down and one may then need to include inequality constraints also on the state variables in order to reconcile the approximate model with the process measurements. As we demonstrate using examples, for a class of problems typically involving bounded disturbances or approximate models, inequality constraints are necessary in order to obtain accurate and physically meaningful state estimates. We focus on this problem.

Satisfying inequality constraints is the domain of mathematical programming. Consequently, any inference process that incorporates constraints is necessarily formulated as a mathematical program. Our interest is in the dynamic estimation formulated as a mathematical program. Our interest is

Correspondence concerning this article should be addressed to J. B. Rawlings.

in the dynamic estimation problem. Hence, our solution employs on-line optimization. Whereas one can view Kalman filtering, among many different alternatives, as an on-line optimization strategy, and our proposed solution reduces to Kalman filtering when we do not consider constraints, the constrained estimation problem cannot escape on-line optimization. While providing the ability to incorporate constraints, on-line optimization introduces practical difficulties. Our proposed solution is moving horizon estimation (MHE). As we demonstrate, MHE bypasses these issues, albeit approximately, and provides, in our opinion, a practical and flexible strategy for constrained state estimation.

Moving-horizon estimation is not a new idea and has been proposed by numerous researchers. Though distinct from control, we find it best from a historical perspective to view MHE as an offshoot of model-predictive control. The success of employing on-line optimization in control as demonstrated by the industrial success of model-predictive control [cf. Qin and Badgwell (1997, 1998)] provided the initial motivation for MHE. The first proposal of unconstrained MHE came from Thomas (1975) and Kwon et al. (1983), although it was Jang et al. (1986) who first proposed unconstrained MHE as an on-line optimization strategy. Many researchers in process systems extended the work of Jang and coworkers. Bequette and coworkers (Bequette, 1991; Ramamurthi et al., 1993) investigated moving-horizon strategies for state estimation as a logical extension of model-predictive control. Kim et al. (1991) and Liebman et al. (1992) investigated moving-horizon strategies for nonlinear data reconciliation. Tjoa and Biegler (1991) and Albuquerque and Biegler (1996, 1997) investigated statistical and numerical issues related to optimization-based nonlinear data reconciliation. Narasimhan and Harikumar (1993a,b) discussed static data-reconciliation strategies incorporating constraints. Marquardt and coworkers (Binder et al., 1998, 2000) discussed multiscale strategies for MHE and the benefits of incorporating constraints in estimation. Bemporad et al. (1999) discussed the application of MHE to hybrid systems. Gesthuisen and Engell (1998) discussed the application of MHE to a pilot-scale polymerization reactor, and Russo and Young (1999) discussed the application of MHE to an industrial polymerization process at the Exxon Chemical Company.

Robertson and Lee (Robertson and Lee, 1995, 2002) and Robertson et al. (1996) investigated the probabilistic interpretation of constraints in estimation. Muske et al. (1993) and Muske and Rawlings (1995) derived some preliminary conditions for the stability of moving-horizon state estimation with inequality constraints. Tyler and Morari (1996) and Tyler (1997) demonstrated how constraints may lead to instability for nonminimum phase systems. Findeisen (1997) investigated the stability and dynamic programming structure of unconstrained, linear MHE with filtering and smoothing updates. Rao and Rawlings (1998) and Rao et al. (1999a) provided sufficient conditions for stability under minimal assumptions in an abstract setting. The theoretical results obtained from those last two articles provide the foundation for this work.

This work provides a complementary development to our previous theoretical results, where we addressed only the issues of existence and stability. Our goal in this article is to develop a general framework for constrained moving-horizon

estimation, with a focus on the practical aspects of the problem. A major focus is on reconciling constraints, particularly those on state variables, with estimation theory, in particular Kalman filtering. As we demonstrate, state constraints alter implicitly the problem structure. The outline of the article is as follows. We begin by introducing the constrained estimation problem in the second section, and then show how moving horizon estimation arises when one considers on-line implementation in the third section. Our focus then shifts and we discuss constraints in the fourth section, in particular the probabilistic interpretation of state constraints and the issue of causality. Using a series of examples of varied complexity, we illustrate the potential utility of incorporating constraints in the inference process. We conclude with a summary of our investigations.

Constrained State Estimation

At time T suppose our observations of the process consist solely of a sequence of discrete measurements $\{y_0, y_1, \dots, y_{T-1}\}$. For simplicity we limit our discussion to the problem where all of the sensors provide measurements simultaneously, though we can extend the proposed strategy *mutatis mutandis* to incorporate multirate sensors. The objective at time T is to reconstruct the evolution of the state of the process $\{x(t); t \geq 0\}$ from the observations $\{y_0, y_1, \dots, y_{T-1}\}$.

We assume we can capture our physical insight of the process with a finite-dimensional (extensions to “infinite-dimensional” or distributed parameter systems are possible, though this problem is far more complex) differential algebraic equation of the form

$$F[x(t), \dot{x}(t), u(t), w(t), t] = 0, \quad (1)$$

where $\dot{x}(\cdot)$ denotes the time derivatives of the state $x(\cdot)$; $u(\cdot)$ denotes measurable exogenous disturbances; and $w(\cdot)$ denotes unmeasurable exogenous disturbances. The disturbance $w(\cdot)$ is typically modeled as a stochastic process and may account also for modeling uncertainty. If we couple our physical insight of the process with the measurements, then we require a model of the process sensors. We relate the observations $y(t)$ to state $x(t)$ using a model of the form

$$y(t) = g[x(t), t] + v(t), \quad (2)$$

where measurement uncertainty is captured in the vector $v(t)$. One commonly assumes the vector $y(t)$ is a normally distributed random variable. We stress that the vector $y(t)$ in Eq. 2 denotes the actual observation, and the vector $v(t)$ denotes the error between the observation $y(t)$ and the predicted sensor reading $g(x(t), t)$.

With the exception of linear and trivial nonlinear process models, we need to discretize the differential algebraic equation (Eq. 1) in order to perform any computation or analysis. At this stage of our discussion, the discretization is conceptual. Discretization is usually performed during optimization. Whether one employs a simultaneous strategy (cf. Biegler, 1997, 1998; Bock et al. 1998), or discretizes first using a DAE solver (cf. Ascher and Pezold 1998) is inconsequential to our discussion, though important when one considers on-line im-

plementation. Hereafter, we suppose that the differential algebraic equation (Eq. 1) is discretized with a zero-order hold on the disturbances $u(\cdot)$ and $w(\cdot)$, yielding the difference equation

$$x_{k+1} = f_d(x_k, u_k, w_k, k), \quad (3)$$

where the integer k denotes the discrete-time index. A typical choice is $t = k\Delta T$, where ΔT denotes the sampling period. The subscripts on the vectors x , u , w , and v denote the value at the points of discretization [for example, $x_k = x(k\Delta T)$]. We assume also the points of discretization (for example, $t_k = k\Delta T$) coincide with the measurement times. Rarely is the equation $f_d(\cdot)$ in Eq. 3 available in algebraic form. Instead, we view the function $f_d(\cdot)$ abstractly as the numerical solution of Eq. 1 with initial condition x_k . The difference equation (Eq. 3) consequently does not include explicitly algebraic constraints, even though the corresponding differential equation (Eq. 1) does.

When we couple physical insight with the process measurements, we need to introduce a measure of uncertainty. The model predictions rarely, if ever, coincide with the process measurements. We need somehow to distribute the errors between the model and sensor measurements. In other words, we need to reconcile our model with the process measurements. Reconciliation in our framework amounts to a trade-off between the vectors w_k and v_k . One may interpret w_k as process disturbances or model uncertainty and the vector v_k as sensor noise. A natural framework to characterize uncertainty is probability theory, where we treat the vectors w_k and v_k as random variables. Our choice of the respective probability distributions provides the reconciliation. A common alternative to probability theory is game theory. In game theory one uses instead, though with often the same result, deterministic uncertainty descriptions of the vectors w_k and v_k (cf. Başar and Bernhard, 1995). Another alternative was proposed recently by Binder et al. (1999). Eschewing both probability and game theory, they view the reconciliation problem instead as the inversion of a compact operator, an ill-posed problem. The trade-off in their framework is the degree of regularization.

When we formulate the state-estimation problem from the perspective of probability theory, we typically model the evolution of the state as a discrete-time Markov process (an equivalent assumption is that the disturbances vectors w_k are independent). As we expect, the process measurements are correlated with the state, and quantity of interest becomes the conditional probability density function of the state evolution $\{x_0, x_1, \dots, x_T\}$ given the process measurements $\{y_0, y_1, \dots, y_{T-1}\}$

$$p(x_0, x_1, \dots, x_T | y_0, y_1, \dots, y_{T-1}). \quad (4)$$

The optimal estimate of the state at time k , given the measurements

$$\{y_0, y_1, \dots, y_{T-1}\},$$

which we denote by $\hat{x}_{k|T-1}$, is then a functional L_T of conditional probability density function (Eq. 4)

$$\begin{aligned} & \{\hat{x}_{0|T-1}, \hat{x}_{1|T-1}, \dots, \hat{x}_{T|T-1}\} \\ & = L_T(p(x_0, x_1, \dots, x_T | y_0, y_1, \dots, y_{T-1})). \end{aligned}$$

A typical choice for the functional L_T is either an expectation or the maximum *a posteriori* Bayesian (MAP) estimate

$$\begin{aligned} & \{\hat{x}_{0|T-1}, \hat{x}_{1|T-1}, \dots, \hat{x}_{T|T-1}\} \in \\ & \arg \max_{\{x_0, x_1, \dots, x_T\}} p(x_0, x_1, \dots, x_T | y_0, y_1, \dots, y_{T-1}). \quad (5) \end{aligned}$$

In this work we focus solely on the Bayesian criterion.

Solving Eq. 5 requires an expression for the conditional probability density function (Eq. 4). Following the developments of Cox (1964) and Jazwinski (1970), we determine the conditional probability density function (Eq. 4) as follows: Using the Markov property, we can express the joint probability of the state as

$$p(x_0, \dots, x_T) = p_{x_0}(x_0) \prod_{k=0}^{T-1} p(x_{k+1} | x_k),$$

where $p_{x_0}(x_0)$ denotes our prior information concerning the initial state of the system. If we assume the measurement noise v_k is independent, then using our model of the sensor (Eq. 2) we have the relationship

$$p(y_0, \dots, y_{T-1} | x_0, \dots, x_{T-1}) = \prod_{k=0}^{T-1} p_{v_k}[y_k - g(x_k, k)].$$

Applying Bayes' rule, we obtain

$$\begin{aligned} & p(x_0, x_1, \dots, x_T | y_0, y_1, \dots, y_{T-1}) \\ & \propto p_{x_0}(x_0) \prod_{k=1}^{T-1} p_{v_k}(y_k - g(x_k, k)) p(x_{k+1} | x_k). \end{aligned}$$

The properties of logarithms allows us to establish the following equality

$$\begin{aligned} & \arg \max_{\{x_0, x_1, \dots, x_T\}} p(x_0, x_1, \dots, x_T | y_0, y_1, \dots, y_{T-1}), \\ & = \arg \max_{\{x_0, x_1, \dots, x_T\}} \log p(x_0, x_1, \dots, x_T | y_0, y_1, \dots, y_{T-1}), \\ & = \arg \max_{\{x_0, x_1, \dots, x_T\}} \sum_{k=0}^{T-1} \log p_{v_k}[y_k - g(x_k, k)] \\ & \quad + \log p(x_{k+1} | x_k) + \log p_{x_0}(x_0). \end{aligned}$$

The last equation is useful, because it allows us to transform the problem (Eq. 5) into a multistage optimization. As we illustrate, compression is conceptually easier to address when the problem structure is multistage.

We have succeeded in transforming the state-estimation problem into a multistage dynamic optimization, though the formulation still requires the specification of the probability density functions. The probability density functions $p_{v_k}(\cdot)$ and

$p_{x_0}(\cdot)$ are commonly chosen as normals. Even though the choice is justified typically by the law of large numbers, one chooses normals, more often than not, because they are mathematically convenient. Evaluating the state transition probability density function $p(x_{k+1}|x_k)$, however, requires the solution of functional difference equation, the discrete-time analog of the Fokker-Planck equation, unless we make the following simplifying assumptions:

A The disturbances w_k are mutually independent;

B $f_d(x_k, u_k, w_k, k) = f(x_k, u_k, k) + w_k$.

(If the vector w_k is a normally distributed random variable, then we can replace assumption **B** with

$$f_d(x_k, u_k, w_k, k) = f(x_k, u_k, k) + Gw_k,$$

where G is a matrix with full column rank. Under these two assumptions, we have

$$p(x_{k+1}|x_k) = p_{w_k}[x_{k+1} - f(x_k, u_k, k)].$$

The probability density function $p_{w_k}(\cdot)$ is also commonly chosen as normal. Assumptions **A** and **B** allow us to cast Eq. 5 as an optimization explicitly in terms of the process model and the probability density functions $p_{v_k}(\cdot)$, $p_{w_k}(\cdot)$, and $p_{x_0}(\cdot)$:

$$\begin{aligned} \arg \max_{\{x_0, x_1, \dots, x_T\}} p(x_0, x_1, \dots, x_T | y_0, \dots, y_{T-1}) \\ = \arg \max_{\{x_0, x_1, \dots, x_T\}} \sum_{k=0}^{T-1} \log p_{v_k}[y_k - g(x_k, k)] \\ + \log p_{w_k}[x_{k+1} - f(x_k, u_k, k)] + \log p_{x_0}(x_0). \end{aligned}$$

If we assume furthermore that the density $p_{x_0}(\cdot)$ is normal with mean \bar{x} and covariance Π_0 , and the densities $p_{w_k}(\cdot)$ and $p_{v_k}(\cdot)$ are normal with zero mean and covariances Q and R , respectively, then we have

$$\begin{aligned} \arg \max_{\{x_0, x_1, \dots, x_T\}} p(x_0, x_1, \dots, x_T | y_0, \dots, y_{T-1}) \\ = \arg \min_{\{x_0, x_1, \dots, x_T\}} \sum_{k=0}^{T-1} \|y_k - g(x_k, k)\|_R^2 \\ + \|x_{k+1} - f(x_k, u_k, k)\|_Q^2 + \|x_0 - \bar{x}\|_{\Pi_0}^2, \end{aligned}$$

where $\|z\|_A^2 = z^T A z$.

The normality assumptions are sufficient for many problems. However, we can improve our descriptions of the random variables w_k , v_k , and x_k by introducing the constraints

$$w_k \in \mathbb{W}_k, \quad v_k \in \mathbb{V}_k, \quad x_k \in \mathbb{X}_k,$$

where the sets \mathbb{W}_k , \mathbb{V}_k , and \mathbb{X}_k are closed and convex. One commonly chooses the sets as polyhedral convex sets, that is,

$$\mathbb{W}_k = \{w_k : w_{\min}^k \leq W_k^k w_k \leq w_{\max}^k\}.$$

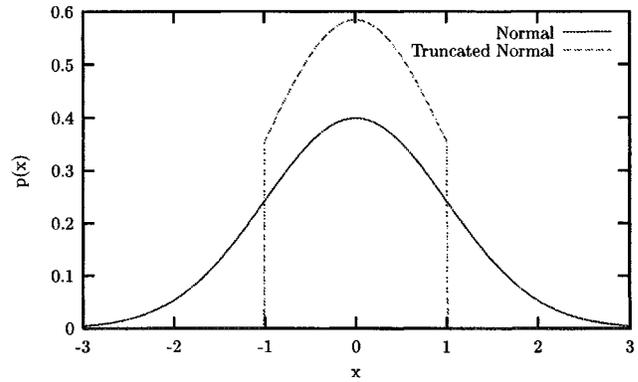


Figure 1. Comparison of a normal and truncated normal-probability density function.

In a probabilistic framework, the constraint sets provide the support for the probability density functions. If, for example, we assume

$$\mathbb{W}_k = \{w_k : -1 \leq w_k \leq 1\},$$

and the probability density function $p_{w_k}(\cdot)$ is a normal with zero-mean and unit variance, then the constraints project the probability density function $p_{w_k}(\cdot)$ onto \mathbb{W}_k , yielding a truncated normal (see Figure 1). One obtains similar results if the probability density function $p_{v_k}(\cdot)$ is coupled with constraints. However, we advise against constraining the vector v_k due to the possibility of outliers. Constraints may amplify the effect of spurious measurements; if one constrains the measurement residual v_k , then the estimate $\hat{x}_{k|T-1}$ may be unable to ignore the spurious measurement, y_k . One can also use constraints to generate asymmetric distributions by piecing together truncated probability density functions as a jigsaw using variable decompositions (Robertson, 1996; Robertson and Lee, 1998).

The probabilistic interpretation and implication of constraints on the state x_k is not as simple. Some of the issues are illustrated in the following simple example. Suppose we have a leaky vessel initially full of a liquid compound A . Let the state x_k denote the mass of A at time k and the vector w_k denote the mass of A that leaks from the vessel during the time interval k to $k + 1$. A simple mass balance yields the model

$$x_{k+1} = x_k + w_k. \tag{6}$$

In addition to the mass balance, we know the state x_k is positive and bounded and the disturbance w_k is negative. One immediate consequence of the state constraint $x_k \geq 0$ is that the state x_k and disturbance w_k are correlated: if the state x_k is small, then the state constraint $x_k \geq 0$ implies that the disturbance w_k is also necessarily small. This result is physically obvious, yet also somewhat surprising. One typically assumes that the exogenous disturbances are independent of the state of the process. If we ignore the effect of recycle and feedback loops, the disturbances are a result of variations in upstream processes unaffected by the state of the downstream process.

Another consequence of state constraints is the violation of causality. If we rewrite the state equation explicitly in terms of the vectors w_k , then we have the equivalent representation

$$x_{k+1} = x_0 + \sum_{j=0}^k w_j. \quad (7)$$

If we suppose that the initial leak w_0 is large, then the future leaks $\{w_1, w_2, \dots\}$ are necessarily small: there is less mass in the vessel that can leak out. Likewise, a large leak at time k requires that past leaks $\{w_0, w_1, \dots, w_j\}$ are small. This causal correlation is equivalent to the correlation between the disturbance vector w_k and the state x_k , because we model the system as a Markov process. Again, one commonly assumes that the disturbances are independent of the state of the process, and in this case they are not. The conclusions from this example are that state constraints can significantly alter the probabilistic structure of the problem. Rarely is this structure explicitly specified in the problem statement, so one should exercise care with state constraints. The advantage of state constraints is that they allow for simplified models: rather than having to develop a detailed correlation between the mass in the vessel x_k and the leak w_k , we can use a simple mass balance in conjunction with constraints. Simplifying the modeling requirements is important because the most time-consuming task in design is model development (Ogunnaike 1995). We discuss the issue of constraints further in the fourth section.

From a system-theoretic perspective, state constraints are nonstandard; one usually chooses an exact model of the plant and, separately, the characteristics of the disturbances, such as boundedness, or that the disturbances are independent and identically distributed with known (zero) mean and variance. The properties of the model and disturbances are distinct. State constraints, on the other hand, implicitly state that the model is in error, because the disturbance-free evolution of the process (Eq. 3), that is, $w_k = 0$, may not automatically satisfy the state constraints for some $x_0 \in \mathbb{X}_0$. Enforcing the state constraints may require a nonzero disturbance sequence $\{w_0, w_1, \dots\}$, that implicitly using the disturbances to account for model error. While not necessarily problematic from a practical standpoint, nowhere in the proposed setup was the issue of “robustness” addressed directly.

Moving-Horizon Estimation

Consider again the problem contained in Eq. 5. Under the assumptions of normality, we can recast the state estimation problem at time T as the following mathematical program ($\{w_k\}_{k=0}^{T-1} \triangleq \{w_0, w_1, \dots, w_{T-1}\}$):

$$\min_{x_0, \{w_k\}_{k=0}^{T-1}} \Phi_T(x_0, \{w_k\}), \quad (8)$$

subject to

$$x_{k+1} = f(x_k, u_k, k) + w_k, \quad (9a)$$

$$y_k = g(x_k, k) + v_k, \quad (9b)$$

$$w_k \in \mathbb{W}_k, \quad x_k \in \mathbb{X}_k, \quad (9c)$$

where

$$\Phi_T(x_0, \{w_k\}) = \sum_{k=0}^{T-1} \|v_k\|_{R^{-1}}^2 + \|w_k\|_{Q^{-1}}^2 + \|x_0 - \bar{x}\|_{\Pi^{-1}}^2.$$

We denote the optimal state and disturbance estimates at time T as the sequences $\{\hat{x}_{k|T-1}\}_{k=0}^T$ and $\{\hat{w}_{k|T-1}\}_{k=0}^{T-1}$. In this formulation, the matrices Q and R are the tuning parameters for reconciling the model with the process measurements. The matrices provide the means by which the errors are distributed between the model and the process sensors. In addition to their statistical significance, the matrices have the following simple interpretation: the matrix Q provides a measure of confidence in the model while the matrix R provides a measure of confidence in the process sensors. Thus, if the matrix Q is “large” relative to R , then we are less confident in the model than in the process sensors, and vice versa. The matrix Π provides a measure of confidence in our knowledge of the initial state: \bar{x} .

Many different options exist for solving the mathematical program (Eqs. 8–9). The problem as formulated requires the solution of a nonlinear program, a computationally demanding, although tractable, problem. If the process model is stiff or has unstable dynamics, a simultaneous strategy, in which the discretization and optimization are performed simultaneously, is often advantageous (Biegler 1997, 1998; Bock et al. 1998). When the process model is linear and the constraints are polyhedral convex sets, the mathematical program reduces to a quadratic program, a far less computationally demanding problem. Regardless of the complexity of the problem, solving the state estimation problem (Eq. 5) on-line is usually impossible, because the size of problem in Eqs. 8–9 grows without bound as we collect more process measurements. On-line implementation therefore requires that we bound the size of the mathematical program in Eqs. 8–9. Consequently, we need a strategy to compress the data. The strategy we employ is *approximate* dynamic programming.

Consider the objective function $\Phi_T(\cdot)$. We can rearrange the objective function $\Phi_T(\cdot)$ by breaking the time interval into two pieces $t_1 = \{k : 0 \leq k \leq T - N - 1\}$ and $t_2 = \{k : T - N \leq k \leq T - 1\}$ as follows:

$$\begin{aligned} \Phi_T(x_0, \{w_k\}_{k=0}^{T-1}) &= \sum_{k=T-N}^{T-1} \|v_k\|_{R^{-1}}^2 + \|w_k\|_{Q^{-1}}^2 \\ &+ \sum_{k=0}^{T-N-1} \|v_k\|_{R^{-1}}^2 + \|w_k\|_{Q^{-1}}^2 + \|x_0 - \bar{x}\|_{\Pi^{-1}}^2. \end{aligned}$$

Because we use a state-variable description of the system (i.e., a Markov process), the quality

$$\sum_{k=T-N}^{T-1} \|v_k\|_{R^{-1}}^2 + \|w_k\|_{Q^{-1}}^2$$

depends only on the state x_{T-N} , disturbance sequence $\{w_k\}_{k=T-N}^{T-1}$, and the process measurements $\{y_k\}_{k=T-N}^{T-1}$. The *principle of optimality* allows us to cast the estimation problem (Eq. 5) as a MHE. Standard dynamic programming argu-

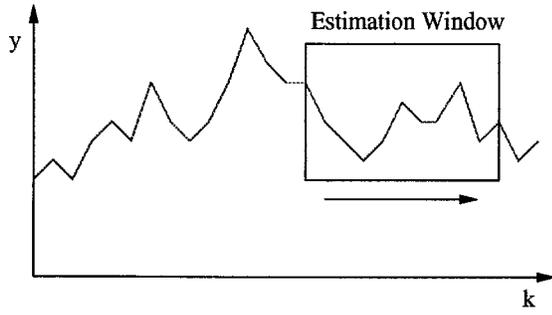


Figure 2. Moving-horizon strategy.

ments allow us to replace the mathematical program in Eqs. 8–9 with the following *equivalent* mathematical program

$$\min_{x_{T-N}, \{w_k\}_{k=T-N}^{T-1}} \sum_{k=T-N}^{T-1} \|v_k\|_{R^{-1}}^2 + \|w_k\|_{Q^{-1}}^2 + Z_{T-N}(x_{T-N}),$$

subject to the constraints in Eq. 9, where

$$Z_\tau(\bar{x}) = \min_{x_0, \{w_k\}_{k=0}^{\tau-1}} \{\Phi_\tau(x_0, \{w_k\}) : x_\tau = \bar{x}\}, \quad (10)$$

subject to the constraints in Eq. 9. The mathematical program (Eq. 10) provides the general structure for MHE. Whereas in the problem (Eqs. 8–9) we considered all of the available process measurements, in MHE we account explicitly only for the last N process measurements. We account for the remaining process measurements using the function $Z_{T-N}(\cdot)$. The name “moving-horizon estimation” arises from the analogy of a sliding-estimation window (see Figure 2).

We refer to the function $Z_\tau(\cdot)$ as the *arrival cost*. Arrival cost is fundamental in estimation, because, by providing a means to compress the data, it allows us to transform the unbounded mathematical problem into an equivalent fixed-dimension mathematical program. The arrival cost compactly summarizes the effect of the data $\{y_k\}_{k=0}^{\tau-1}$ on the state x_τ , thereby allowing us to fix the dimension of the optimization. We can view arrival cost as the analogy of the “cost to go” in standard backward dynamic programming. In probabilistic terms, the arrival cost generates the conditional density function $p(x_\tau|y_0, \dots, y_{\tau-1})$, and vice versa: the arrival cost is proportional to the negative logarithm of the conditional density function $p(x_\tau|y_0, \dots, y_{\tau-1})$. Hence, we can view arrival cost as an equivalent statistic (Striebel, 1965) for the conditional density function $p(x_\tau|y_0, \dots, y_{\tau-1})$. Further discussion on the properties of arrival cost can be found in Rao and Rawlings (1998) and Rao (2000).

Arrival cost provides a general method for compressing the process data. An excellent example of arrival cost is the Riccati equation arising in Kalman filtering. Consider the problem in Eqs. 8–9, where we assume the model is linear

$$x_{k+1} = Ax_k + w_k, \quad y_k = Cx_k + v_k$$

and ignore the constraints \mathbb{X}_k and \mathbb{W}_k . If we use the Kalman filter covariance update formula

$$\begin{aligned} \Pi_T &= GQG^T \\ &+ A\Pi_{T-1}A^T - A\Pi_{T-1}C^T(R - C\Pi_{T-1}C^T)^{-1}C\Pi_{T-1}A^T \end{aligned} \quad (11)$$

subject to the initial condition $\Pi_0 = \Pi$, then, assuming the matrix Π_T is invertible, we can express the arrival cost explicitly as

$$Z_T(x) = (x - \hat{x}_T)^T \Pi_T^{-1} (x - \hat{x}_T) + \Phi_T^*,$$

where the \hat{x}_T denotes the optimal estimate at time T given the measurements $\{y_k\}_{k=0}^{T-1}$ and Φ_T^* denotes the optimal cost at time T . From the preceding arguments, we have

$$\begin{aligned} \min_{x_0, \{w_k\}_{k=0}^{T-1}} \phi_T(x_0, \{w_k\}) \\ \equiv \min_{x_{T-N}, \{w_k\}_{k=T-N}^{T-1}} \sum_{k=T-N}^{T-1} \|v_k\|_{R^{-1}}^2 + \|w_k\|_{Q^{-1}}^2 \\ + \|x_{T-N} - \hat{x}_{T-N}\|_{\Pi_{T-N}^{-1}}^2 + \Phi_{T-N}^*. \end{aligned}$$

We can extract the Kalman filter by considering a horizon of $N = 1$. For this scenario, we have

$$\begin{aligned} \Phi_T(x_{T-1}, w_{T-1}) &= v_{T-1}^T R^{-1} v_{T-1} + w_{T-1}^T Q^{-1} w_{T-1} \\ &+ (x_{T-1} - \hat{x}_{T-1})^T \Pi_{T-1}^{-1} (x_{T-1} - \hat{x}_{T-1}). \end{aligned}$$

Substituting in the model equations, evaluating the minimum with respect to w_{T-1} and x_{T-1} , and using some algebra, we obtain the well-known result

$$\hat{x}_T = A\hat{x}_{T-1} + L(y_T + CA\hat{x}_{T-1})$$

for the Kalman filter, where

$$L = A\Pi_{T-1}C^T(R + C\Pi_{T-1}C^T)^{-1}.$$

Unfortunately, algebraic expressions for arrival cost do not exist when either constraints are present or the process model is nonlinear. As these are the problems of interest, we need to generate *approximate* algebraic expressions for the arrival cost. At one extreme, we can discard the past information by approximating the arrival cost as a constant function. At the other extreme, we can ignore the current measurements and consider only the past measurements by approximating the arrival cost with the extended real-valued function

$$\hat{Z}_\tau(x_\tau) = \begin{cases} \Phi_\tau^* & x_\tau = \hat{x}_\tau \\ \infty & x_\tau \neq \hat{x}_\tau. \end{cases}$$

Both of these choices are undesirable. Rarely are we completely ignorant or informed of the value of the state x_τ . One strategy to approximate the arrival cost is to use a first-order Taylor series approximation of the model around the esti-

mated trajectory $\{\hat{x}_k\}_{k=0}^T$. This strategy approximates the arrival cost with an extended Kalman filter covariance update formula. We interpret this strategy as a neighboring extremal paths strategy in the context of estimation. Neighboring extremal paths are used to generate approximate optimal feedback laws for nonlinear systems by employing an extended linearization (Bryson and Ho, 1975). The basic idea is as follows: If the deviation from the optimal path is small, then a linear approximation at the optimal path accurately describes the neighboring path.

If we let

$$A_k = \left. \frac{\partial f(x_k, u_k, k)}{\partial x_k} \right|_{\hat{x}_T}, \quad C_k = \left. \frac{\partial g(x_k)}{\partial x_k} \right|_{\hat{x}_T},$$

then we obtain the extended Kalman filter covariance recursively from the equation

$$\Pi_{T+1} = Q + A_T \left(\Pi_T - \Pi_T C_T^T (R + C_T \Pi_T C_T^T)^{-1} C_T \Pi_T \right) A_T^T, \quad (12)$$

subject to the initial condition $\Pi_0 = \Pi$. The choice

$$\hat{Z}_T(\bar{x}) = \|\bar{x} - \hat{x}_T\|_{\Pi_T}^2 \quad (13)$$

summarizes our best available knowledge, to a first-order approximation, without introducing extra knowledge not available from the measurements. Using the extended Kalman filter to approximate the arrival cost has many advantages. When there are no constraints, one can view the estimator as an iterated extended Kalman filter. When the process model is linear, the estimator reduces to a Kalman filter.

One needs to be wary of divergence (instability) when approximating the arrival cost. So long as the approximate arrival cost $\hat{Z}_T(\cdot)$ satisfies certain technical conditions, one is guaranteed nondivergence, or stability (Rao and Rawlings, 1998). When the process model is linear, the Kalman filter covariance, regardless of whether there are constraints, yields a stable estimator (Rao et al., 1999b). However, when the process model is nonlinear, the extended Kalman filter covariance does not guarantee stability, and additional measures are necessary to guarantee stability. In practical terms, there should be a degree of forgetting: the estimator should not weigh the past data too heavily. One property of the Kalman filter is that it exponentially forgets the past data (cf. Anderson, 1999). If one is concerned about estimator divergence, then adding a “forgetting factor” to the approximate arrival cost improves the estimator’s “robustness.” A simple strategy for generating a forgetting factor is to premultiply the approximate arrival cost by a scalar $\alpha \in (0, 1)$:

$$\hat{Z}_T(\bar{x}) = \alpha \|\bar{x} - \hat{x}_T\|_{\Pi_T}^2.$$

The interested reader is referred to Rao (2000) for further discussion regarding forgetting factors in constrained moving-horizon estimation.

We therefore formulate MHE at time T as the solution to the following mathematical program

$$\min_{x_{T-N}, \{w_k\}_{k=T-N}^{T-1}} \hat{\Phi}_T(x_{T-N}, \{w_k\}),$$

subject to the constraints

$$\begin{aligned} x_{k+1} &= f(x_k, u_k, k) + w_k, \\ y_k &= g(x_k, k) + v_k, \\ w_k &\in \mathbb{W}_k, \quad x_k \in \mathbb{X}_k, \end{aligned}$$

where

$$\begin{aligned} \hat{\Phi}_T(x_{T-N}, \{w_k\}) \\ = \sum_{k=T-N}^{T-1} \|w_k\|_Q^2 + \|v_k\|_R^2 + \|x_{T-N} - \hat{x}_{T-N}\|_{\Pi_{T-N}}^2. \end{aligned}$$

We denote, with abuse of notation, the optimal state and disturbance estimates at time T as the sequences $\{\hat{x}_{k|T-1}\}_{k=T-N}^T$ and $\{\hat{w}_{k|T-1}\}_{k=T-N}^{T-1}$. Unlike the “full information” problem (Eqs. 8–9), the MHE estimator generates only truncated estimates—the consequence of considering only the data sequence $\{y_k\}_{k=T-N}^{T-1}$. The pair $(\hat{x}_{T-N}, \Pi_{T-N})$ summarizes the prior information at time $T-N$. The vector \hat{x}_{T-N} is the moving-horizon state estimate at time $T-N$ and the matrix Π_{T-N} is the solution to Eq. 12 subject to the initial condition Π_0 . When $T \leq N$, MHE is equivalent to the full information estimator. For simplicity, let $\hat{x}_T \triangleq \hat{x}_{T|T-1}$. This formulation of MHE was first proposed by Muske et al. (1993) and Robertson et al. (1996).

The choice of the horizon length N is a tuning parameter in MHE. The performance of MHE improves as one increases the horizon length, though with diminishing returns once N is sufficiently large. However, the computational cost also increases with the horizon length. One needs to reconcile these two objectives when choosing the horizon length. From the theoretical standpoint, MHE is stable so long as the horizon length is greater than the order, or the observability index, of the system. A practical rule of thumb is to choose the horizon length as twice the order of the system.

Constraints

The strength of MHE is the ability to incorporate constraints in estimation. One can plausibly argue that nonlinear dynamics also motivate the use of MHE. However, we believe there are many competitive alternatives to *unconstrained* MHE. From a theoretical standpoint, one strength of MHE is that it provides stability guarantees (Rao and Rawlings, 1998). However, many other state estimation strategies also provide stability guarantees. For example, one can also construct a stable estimator using a local coordinate transformation by output injection (Bestle and Zeitz, 1983; Krener and Isidori, 1983). We note that, unlike differential geometric methods, moving horizon strategies are applicable to a larger class of problems. In particular, any feedback lineariz-

able system also can be stabilized with a moving-horizon controller (Meadows et al., 1995). We expect a dual result holds for estimation.

Stability guarantees are important, but performance is the predominant concern. The extended Kalman filter provides only weak local stability guarantees (cf. Song and Grizzle, 1995), yet is the *de facto* choice for estimating the state with nonlinear process models. We can view unconstrained MHE as a form of extended Kalman filtering or, rather, the extended Kalman filter as a form of unconstrained MHE. The difference between the two strategies is the degree of optimization: the extended Kalman filter takes only one Newton step, while unconstrained MHE takes as many Newton steps as necessary to satisfy the (local) optimality conditions. We therefore view unconstrained MHE as a form of iterated extended Kalman filtering and the extended Kalman filter as a suboptimal strategy for unconstrained MHE with a horizon length $N=1$. One reason for the success of the extended Kalman filter is that often most of the cost reduction in optimization is obtained during the first few Newton steps. Performance rarely improves tangibly even if one iterates further.

Without constraints, MHE often tends to perform the same as the extended or, rather, the iterated extended Kalman filter. Performance changes when one adds constraints to the problem. Constraints therefore motivate the use of MHE. We can best illustrate the potential of MHE with the following examples.

Example of inequality constraints yielding improved estimates

Consider the following discrete-time system (this state-space system is a realization of the following system: $G(s) =$

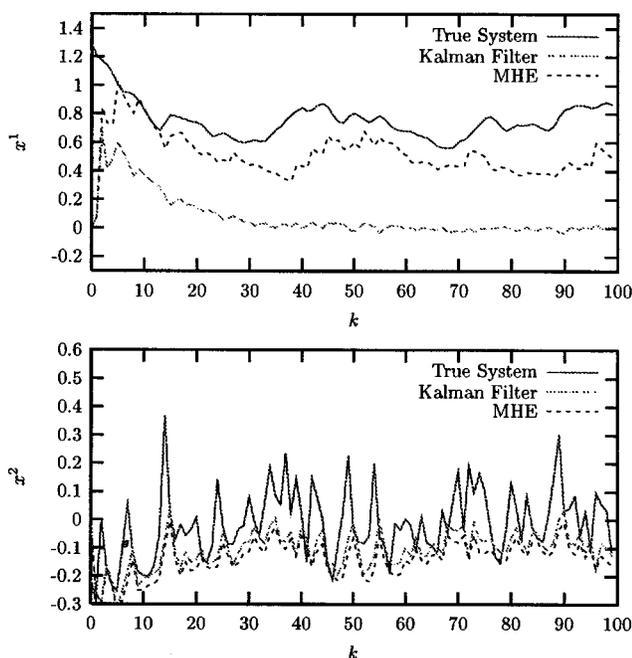


Figure 3. Comparison of estimators for Example 4.1.

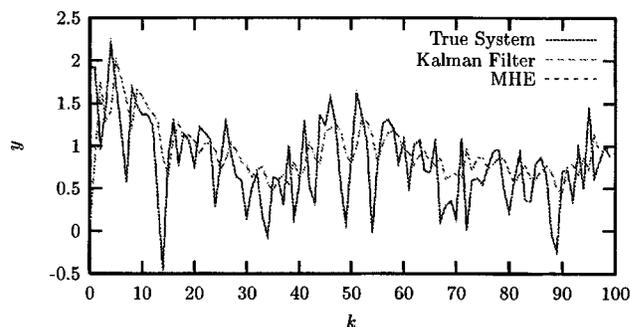


Figure 4. Comparison of output predictions for Example 4.1.

$(-3s+1)/(s^2+3s+1)$ sampled with a zero-order hold and sampling time of 0.3)

$$x_{k+1} = \begin{bmatrix} 0.9962 & 0.1949 \\ -0.1949 & 0.3815 \end{bmatrix} x_k + \begin{bmatrix} 0.03393 \\ 0.1949 \end{bmatrix} w_k,$$

$$y_k = [1 \quad -3]x_k + v_k. \quad (14)$$

We assume $\{v_k\}$ is a sequence of independent, zero-mean, normally distributed random variables with covariance 0.01, and $w_k = |z_k|$ where $\{z_k\}$ is a sequence of independent, zero-mean, normally distributed random variables with unit covariance. We also assume the initial state x_0 is normally distributed with zero mean and covariance equal to the identity.

We formulate the constrained estimation problem with $Q = 1$, $R = 0.01$, $\Pi_0 = 1$, and $\bar{x} = 0$. For the MHE, we choose $N = 10$. To capture our knowledge of the random sequence w_k , we add the inequality constraint $w_k \geq 0$. Note, this formulation yields the *optimal* Bayesian estimate. A comparison of the Kalman filter, full information estimator, and MHE for a single realization of Eq. 14 is shown in Figure 3–4. As expected, the performance of the constrained estimators is superior to the Kalman filter, because the constrained estimators possess, with the addition of the inequality constraints, the proper statistics of the disturbance sequence, w_k . Hence, the constrained estimation problem formulated earlier accurately models the random variable w_k .

If we consider the statistics of the random variable w_k , it is important to note that the mean is not zero and the covariance is not one. Rather, the mean is $2/\sqrt{2\pi}$ and the covariance is $(1 - 2/\pi)$. When we consider the negative inverse logarithm of the probability density function, however, we have

$$-\log p_{w_k}(w_k) \propto \frac{1}{2} w_k' w_k \quad \text{for } w_k \geq 0.$$

Note, therefore, that constraints allow for non-Gaussian distributions.

Leak detection and inventory estimation

Consider the problem of detecting the location and magnitude of a leak in the wastewater treatment process shown in Figure 5. We suppose the process is described by the follow-

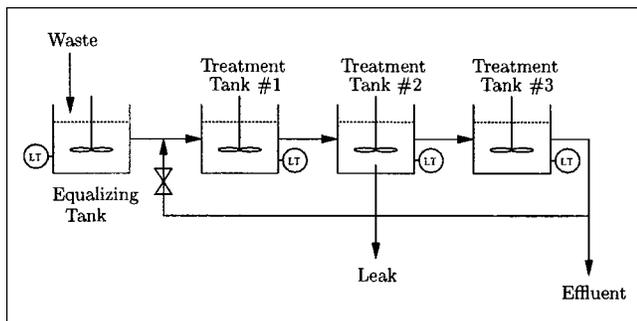


Figure 5. Tank process.

ing linear state-space model

$$x_{k+1} = \begin{bmatrix} 0.89168 & 0 & 0 & 0 & 1.0 \\ 0.10832 & 0.90518 & 0 & 0.04306 & 0 \\ 0 & 0.09482 & 0.89524 & 0 & 0 \\ 0 & 0 & 0.10476 & 0.89235 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} x_k + \begin{bmatrix} -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} w_k, \quad y_k = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & m \end{bmatrix} x_k + v_k.$$

The physical meaning of the entries in state variable x_k are given in Table 1. We choose $m = 1$ in the y_k equation just given when we suppose the mass of waste entering the process is measured, and $m = 0$ otherwise. We suppose the mass of each tank and the mass flow rate of waste entering the process are measured with error covariance

$$R = \text{diag}[8 \ 8 \ 8 \ 8 \ 4].$$

Table 1. State Description for Example

$x^{(1)}$	Mass in equalizing tank
$x^{(2)}$	Mass in Tank No. 1
$x^{(3)}$	Mass in Tank No. 2
$x^{(4)}$	Mass in Tank No. 3
$x^{(5)}$	Mass of waste entering equalizing tank

As the leak is limited to waste tank No. 2, the process was simulated with $w_k = |z_k|$, where z_k is a normally distributed random variable with covariance matrix

$$Q_z = \text{diag}[0 \ 0 \ 5 \ 0 \ 15].$$

As the location of the leak is unknown (to the estimator), we design the estimator with the covariance matrix

$$Q = \text{diag}[5 \ 5 \ 5 \ 5 \ 15].$$

We furthermore added the constraints $w_k \geq 0$ and $x_k \geq 0$ in order to satisfy the mass balances: mass is only lost through a leak, and the tanks must have positive mass. A horizon of $N = 10$ was chosen.

Two separate scenarios were considered (flow measured and unmeasured) along with a control where there is no leak. The results of the simulations are shown in Table 2. As one would expect, both the Kalman filter and MHE are able to detect the leak. The ability to detect the leak degrades when the flow rate is unmeasured. This result is expected, as less information is available to both estimators. The benefit of constraints arise when one attempts to estimate the total losses. While MHE is able to provide a fairly accurate estimate of the total losses, the Kalman filter underestimates the total losses. The Kalman filter also provides *negative* estimates for the losses in the equalizing tank and tank No. 1 in all four scenarios. Furthermore, when there is no leak, the Kalman filter predicts a net addition of mass to the tank system, which obviously is physically impossible. One can attribute this difference to the addition of constraints; the only difference between the two algorithms.

Note also that the constrained estimates are slightly biased away from zero in the tanks not leaking. Recall from the pre-

Table 2. Simulation Results of Example

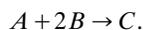
Scenario		Total Losses	Mean Losses (by Tank)			
			Equal.	No. 1	No. 2	No. 3
Flow Measured	Actual	948.08	0	0	1.8962	0
	MHE	992.82	0.2429	0.2387	1.1253	0.3176
	KF	593.77	-0.2538	-0.0610	1.0262	0.0425
No leak	Actual	0	0	0	0	0
	MHE	295.49	0.2418	0.2698	0.2925	0.27001
	KF	-186.44	-0.2552	-0.0129	0.0113	0.0014
Flow Unmeasured	Actual	916.81	0	0	1.8336	0
	MHE	907.54	0.1074	0.2427	1.0664	0.3123
	KF	405.24	-0.5722	-0.0626	0.9761	0.03860
No leak	Actual	0	0	0	0	0
	MHE	244.87	0.1655	0.2729	0.2794	0.2648
	KF	-335.74	0.5732	-0.0169	-0.0032	-0.0010

vious example that a truncated normal does not have a zero mean. Rather, the mean is $2\sigma/\sqrt{2\pi}$, where σ is the standard deviation of the corresponding normally distributed random variable. Any automated procedure involving hypothesis testing needs to account for this fact. This point is clearly illustrated when we simulate the waste treatment process without any leaks. When the flow rate is measured, the constrained estimate is worse than the unconstrained estimate. The poor estimates are due to the positive mean values: the constrained estimates have a mean bias of roughly 0.25. If we remove the bias, the estimate for the total leak is roughly zero, as desired. However, if we remove the bias from the simulation where there is a leak in Tank No. 2, then the estimate for the total leak is the same as the Kalman filter. This example illustrates some of the issues one needs to be wary of when implementing constraints. While the constrained estimators provide a good estimate of the total losses when there is a leak, MHE and the Kalman filter both provide poor estimates when there are no leaks. The problem stems from an incorrect model of the process: the true process has no leaks, while the model assumes a leak in each tank. Nevertheless, one would normally use such a model in fault detection. Hence, any analysis would need to account for this discrepancy.

A “proper” strategy is to formulate this problem as a constrained signal-detection problem. One would model all leak possibilities and then discriminate between the various scenarios using hypothesis testing. An alternative is to employ mixed-integer programming (cf. Gatzke and Doyle, 1999). As the focus of this article is not fault detection, but rather constrained monitoring, we do not pursue this topic further.

Semibatch Reactor

Consider the stirred-tank reactor depicted in Figure 6 where the following liquid-phase exothermic reaction occurs



The state estimation problem, inspired by the problem considered in Rawlings et al. (1989), is to estimate precisely the concentration of A in the reactor. Because overaddition of B leads to product degradation, precise concentration estimates of A as a function of time are necessary to complete the reaction without overaddition of B . We suppose only temperature measurements corrupted with sensor noise are available. Furthermore, we suppose the exact reaction kinetics are unknown with the exception of the heat of reaction ΔH_r . The extent of the reaction is estimated using reaction calorimetry (cf. Schuler and Schmidt, 1992).

Under standard assumptions, such as negligible potential and kinetic energy effects, constant density, uniformly homogeneous mixture, and no phase transition, we simulated the reactor using the following model:

$$\dot{V} = F,$$

$$\dot{A} = -k_0 \exp\left(-\frac{E}{T}\right) AB^2 - \frac{F}{V} A,$$

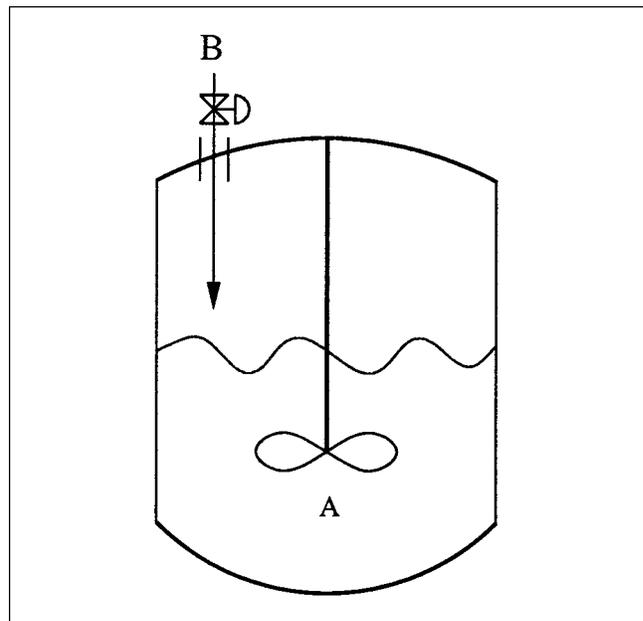


Figure 6. Reactor.

$$\dot{B} = -2k_0 \exp\left(-\frac{E}{T}\right) AB^2 + \frac{F}{V} (C_{Bf} - B),$$

$$\dot{T} = -\frac{\Delta H_r}{\rho C_p} k_0 \exp\left(-\frac{E}{T}\right) AB^2 + \frac{F}{V} (T_f - T)$$

$$+ \frac{UA}{\rho C_p FV} (T_c - T).$$

The model parameters are listed in Table 3. The flow-rate profile, though scaled differently, is the one used in the operation of the industrial reactor described by Rawlings et al. (1989). To account for imperfect cooling and modeling inaccuracies, we assumed the cooling-water temperature fluctuates. The flow-rate profile and the cooling-water temperature used in the simulation are shown in Figure 7. We suppose the temperature measurements are available every 30 s, corrupted with zero mean and unit variance Gaussian noise. The measured and actual reactor temperature are shown in Figure 8.

The estimator has available only the following simplified time-varying linear model based on reaction calorimetry (we also considered a model where the cooling water tempera-

Table 3. Parameters for Example

k_0	$9 \times 10^{11} \text{ mol}^{-2} \cdot \text{min}^{-1}$	$V(0)$	100 L
E	6,000	$A(0)$	0.5 mol/L
ρ	1000 g/L	$B(0)$	0 mol
C_p	0.239 J/g·K	$T(0)$	300 K
UA	$2 \times 10^5 \text{ J/min} \cdot \text{K}$		
T_f	300 K		
T_c	300 K		
C_{Bf}	2.2 mol/L		
$-\Delta H_r$	$5 \times 10^4 \text{ J/mol}$		

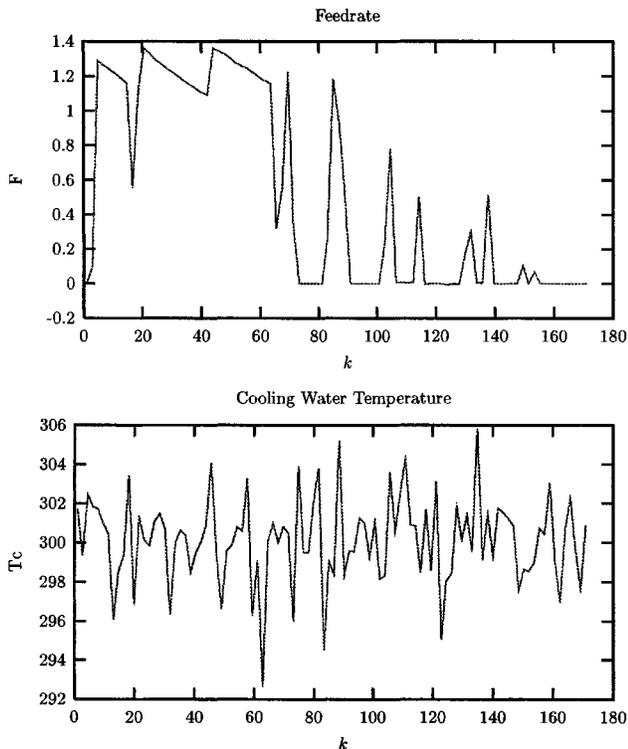


Figure 7. Reactor inputs.

ture fluctuations were included as a second disturbance: our simulation results were no different)

$$\begin{aligned} \dot{V} &= F, \\ \dot{A} &= r - \frac{F}{V}A, \\ \dot{B}(t) &= 2r + \frac{F}{V}(C_{Bf} - B), \\ \dot{T} &= \frac{\Delta H_r}{\rho C_p} r + \frac{F}{V}(T_f - T) + \frac{UA}{\rho C_p V}(T_c - T) \\ dQ_r &= dw. \end{aligned}$$

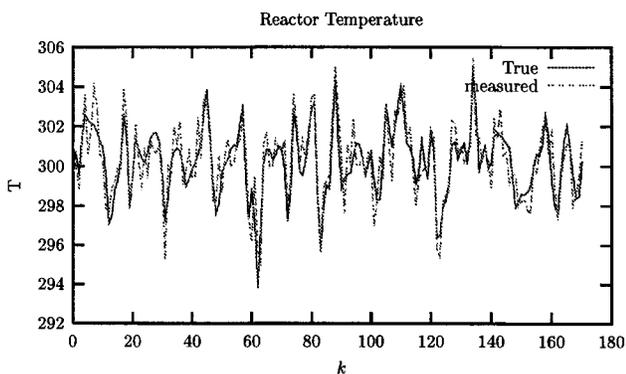


Figure 8. Measured and actual reactor temperature.

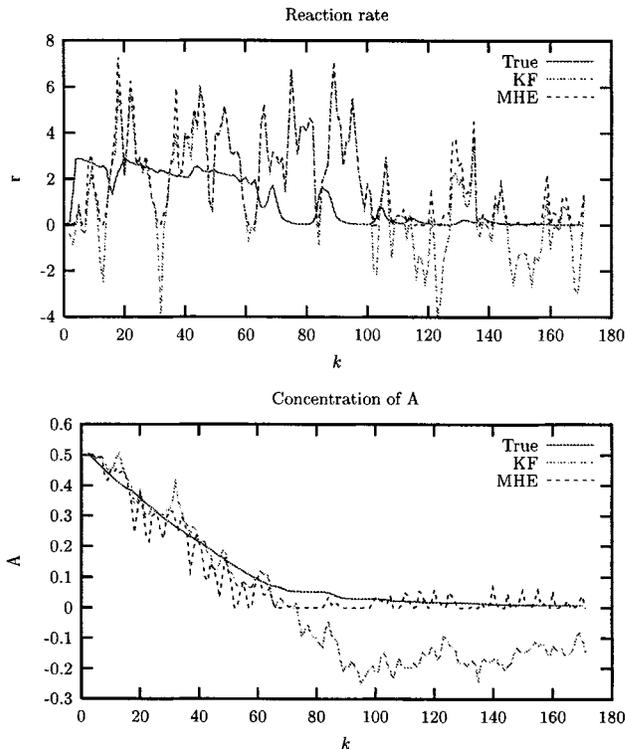


Figure 9. Comparison of estimates.

The trick in reaction calorimetry is to estimate the reaction rate $r(\cdot)$ from the energy balance. The model was discretized with a zero-order hold and a sampling period of 30 s. The horizon length was $N = 10$.

The advantage of the simplified model is that the reaction kinetics need not be known. However, as pointed out by DeVallière and Bonvin (1990) and M'hamdi, et al. (1996), spurious estimates may result due to negative estimates of the reaction rate. We therefore constrain both the reaction rate and concentrations to be positive. We tuned the estimator with $Q = I$ and $\Pi_0 = I$ and initialized the estimator with the "true" initial conditions.

The results of the simulation are shown in Figure 9. Both the Kalman filter and MHE overestimate the actual reaction rate. This mismatch is due to fluctuations in the cooling-water temperature. The addition of the constraints prevents MHE from estimating negative reaction rates and negative concentrations of A . Because MHE does not estimate negative reaction rates, the MHE estimate of the reaction rate is larger than the Kalman filter estimate. Consequently, without the constraint on the concentrations, MHE also would estimate negative concentrations of A . The reason that the estimates are positive, even though the estimate of the reaction rate is too large, is due to smoothing. At each sampling time, MHE semi-implicitly estimates the entire reaction rate and concentration profile. We refer to these estimates as the smoothed estimates ($\hat{x}_{k|T}$ for $k \leq T$). The results shown in Figure 9 are only the tail of the estimated trajectory ($\hat{x}_{T|T-1}$) and need not mutually satisfy the energy and mass balances. The smoothed estimates, however, mutually satisfy the energy and mass balances.

Conclusions

We have discussed MHE in the context of constrained process monitoring. MHE, as we have demonstrated through examples, is a practical and powerful strategy for constrained process monitoring. MHE allows the use of additional physical knowledge about systems, such as constraints and nonlinear dynamics, unavailable with other methods. While the ability to incorporate nonlinear dynamics is important, the distinguishing feature of MHE is the ability to incorporate inequality constraints. One can show, in particular, that MHE reduces to a Kalman filter or iterated extended Kalman filter when constraints are not present. Hence, we can view MHE as an extension of Kalman filtering.

Inequality constraints arise in many different contexts. We have illustrated the importance of inequality constraints in the following situations.

Truncated Distributions. One often possesses prior knowledge in the form of bounds on the disturbances, state variables, and unknown parameters. If we consider the leak detection example, the leaks and tank volumes are always positive. Failure to incorporate this information in the estimator, as illustrated in the inequality constraints and leak detection examples, may lead to poor estimates.

Asymmetric Distributions. By piecing together truncated distributions, it is possible to generate asymmetric distributions. The need for asymmetric distributions is illustrated in the leak detection example, where mass enters the equalizing tank at a different frequency and magnitude than it leaves. The inability to model this behavior can lead to spurious estimates, as illustrated by the Kalman filter's low estimate of the total losses due to the leak.

Model Simplification. Whereas truncated and asymmetric distributions only alter the description of the unknown disturbances, state constraints alter the probabilistic structure of the estimation problem by correlating the disturbances with the state. The advantage is that one can use the correlations to simplify the model significantly. This idea is illustrated in the semibatch reactor example, where a simplified model of the semibatch reactor using reaction calorimetry coupled with constraints allows for accurate concentration estimates.

Reconciling Conservation Laws. Poor measurements can lead to estimates that violate the conservation laws used to model the system. As one often expects the estimates to satisfy the conservation laws, direct enforcement may require inequality constraints. In the semibatch reactor example, the estimates of the reaction rate are too high, and the estimates need to be adjusted in order to prevent negative concentration estimates. From a numerical perspective, one can use constraints to prevent the optimization algorithm from choosing spurious iterates that lead to computational problems regarding the solution of the conservation laws and the associated constitutive relations.

The strength and weakness of MHE is the use of mathematical programming. For reasonable models, the optimization problems can be solved in a few seconds on desktop computers using standard software. However, for some problems this performance is insufficient. With the increasing power of computers and improved algorithms (that is, algorithms now solve quadratic programs in polynomial time),

MHE will become an alternative for an expanding class of constrained process monitoring problems in the near future.

Acknowledgments

The authors gratefully acknowledge the financial support of the industrial members of the Texas-Wisconsin Modeling and Control Consortium and NSF through Grant No. CTS-9708497. All simulations were performed using Octave (<http://www.octave.org>). Octave is freely distributed under the terms of the GNU General Public License.

Literature Cited

- Albuquerque, J., and L. T. Biegler, "Data Reconciliation and Gross-Error Detection for Dynamic Systems," *AIChE J.*, **42**, 2841 (1996).
- Albuquerque, J., and Biegler, L. T. "Decomposition Algorithms for On-Line Estimation with Nonlinear DAE Models," *Comput. Chem. Eng.*, **21**, 283 (1997).
- Anderson, B. D. O., "From Wiener to Hidden Markov Models," *IEEE Control Syst. Mag.* p. 41 (1999).
- Ascher, U. N., and L. R. Petzold, *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*, SIAM, Philadelphia (1998).
- Başar, T., and P. Bernhard, *H[∞]-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, Birkhäuser, Boston, (1995).
- Bemporad, A., D. Mignone, and M. Morari, "Moving Horizon Estimation for Hybrid Systems and Fault Detection," *Proc. of 1999 American Control Conf.*, San Diego, CA (1999).
- Bequette, B. W., "Nonlinear Predictive Control Using Multi-Rate Sampling," *Can. J. Chem. Eng.*, **69**, 136 (1991).
- Bestle, D., and M. Zeitz, "Canonical Form Observer Design for Non-Linear Time-Variable Systems," *Int. J. Control*, **38**(2), 419 (1983).
- Biegler, L. T., "Advances in Nonlinear Programming Concepts for Process Control," *IFAC Adchem '97: Int. Symp. on Advanced Control of Chemical Processes*, Banff, Alta., Canada, p. 587 (1997).
- Biegler, L. T., "Efficient Solution of Dynamic Optimization and NMPC Problems," *Int. Symp. on Nonlinear Model Predictive Control*, Ascona, Switzerland, (1998).
- Binder, T., L. Blank, W. Dahmen, and W. Marquardt, "Towards Multiscale Dynamic Data Reconciliation," *NATO ASI on Nonlinear Model Based Process Control*, Kluwer, Dordrecht, The Netherlands (1998).
- Binder, T., L. Blank, W. Dahmen, and W. Marquardt, "Regularization of Dynamic Data Reconciliation Problems by Projection," *ADCHEM*, 2000, Pisa, Italy, p. 689 (2000).
- Bock, H. G., M. Diehl, D. B. Leineweber, J. P. Schlöser, "A Direct Multiple Shooting Method for Real-Time Optimization of Nonlinear DAE Processes," *Int. Symp. on Nonlinear Model Predictive Control*, Ascona, Switzerland (1998).
- Bryson, A. E., and Y. Ho, *Applied Optimal Control*, Hemisphere, New York (1975).
- Cox, H., "On the Estimation of State Variables and Parameters for Noisy Dynamic Systems," *IEEE Trans. Automat. Cont.*, **AC-9**(1), 5 (1964).
- DeVallière, P., and D. Bonvin, "Application of Estimation Techniques to Batch Reactors: III. Modeling Refinements Which Improve the Quality of State and Parameter Estimation," *Comput. Chem. Eng.*, **14**, 799 (1990).
- Findeisen, P., "Moving Horizon State Estimation of Discrete Time Systems," Master's Thesis, Univ. of Wisconsin-Madison (1997).
- Gatzke, E. P., and F. J. Doyle III, "Moving Horizon Estimation and Control of an Experimental Process," *AIChE Meeting*, Dallas, TX (1999).
- Gesthuisen, R., and S. Engell, "Determination of the Mass Transport in the Polycondensation of Polyethyleneterephthalate by Nonlinear Estimation Techniques," *Proc. 1998 IFAC DYCOPS Symp.*, Corfu, Greece (1998).
- Jang, S-S., B. Joseph, and H. Mukai, "Comparison of Two Approaches to On-Line Parameter and State Estimation of Nonlinear Systems," *Ind. Eng. Chem. Proc. Des. Dev.*, **25**, 809 (1986).

- Jazwinski, A. H., *Stochastic Processes and Filtering Theory*, Academic Press, New York (1970).
- Kailath, T., "A View of Three Decades of Linear Filtering Theory," *IEEE Trans. Inform. Theory*, **IT-20**, 146 (1974).
- Kim, I., M. Liebman, and T. Edgar, "A Sequential Error-in-Variables Method for Nonlinear Dynamic Systems," *Comput. Chem. Eng.*, **15**, 663 (1991).
- Krener, A. J., and A. Isidori, "Linearization by Output Injection and Nonlinear Observers," *Syst. Control Lett.*, **47** (1983).
- Kwon, W. H., A. M. Bruckstein, and T. Kailath, "Stabilizing State-Feedback Design via the Moving Horizon Method," *Int. J. Control*, **37**, 631 (1983).
- Liebman, M., T. Edgar, L. Lasdon, "Efficient Data Reconciliation and Estimation for Dynamic Processes Using Nonlinear Programming Techniques," *Comput. Chem. Eng.*, **16**, 963 (1992).
- Meadows, E. S., M. A. Henson, J. W. Eaton, and J. B. Rawlings, "Receding Horizon Control and Discontinuous State Feedback Stabilization," *Int. J. Control*, **62**, 1217 (1995).
- M'hamdi, A., A. Helbig, O. Abel, and W. Marquardt, "Newton-Type Receding Horizon Control and State Estimation," *Proc. 1996 IFAC World Congress*, San Francisco, CA, p. 121 (1996).
- Muske, K. R., and J. B. Rawlings, "Nonlinear Moving Horizon State Estimation," *Methods of Model Based Process Control, Nato Advanced Study Institute Series: E Applied Sciences 293*, R. Berber, ed., Kluwer, Dordrecht, The Netherlands, p. 349 (1995).
- Muske, K. R., J. B. Rawlings, and J. H. Lee, "Receding Horizon Recursive State Estimation," *Proc. 1993 American Control Conf.*, p. 900 (1993).
- Narasimhan, S., and P. Harikumar, "A Method to Incorporate Bounds in Data Reconciliation and Gross Error Detection: i. the Bounded Data Reconciliation Problem," *Comput. Chem. Eng.*, **17**, 1115 (1993a).
- Narasimhan, S., and P. Harikumar, "A Method to Incorporate Bounds in Data Reconciliation and Gross Error Detection—ii. Gross Error Detection Strategies," *Comput. Chem. Eng.*, **17**, 1121 (1993b).
- Ogunnaike, B. A., "A Contemporary Industrial Perspective on Process Control Theory and Practice," *Proc. IFAC Symp. Dynamics and Control of Chemical Reactors, Distillation Columns and Batch Processes*, Helsingor, Denmark, p. 1 (1995).
- Qin, J. S., and T. A. Badgwell, "An Overview of Nonlinear Model Predictive Control Applications," *Symp. on Nonlinear Model Predictive Control*, Ascona, Switzerland (1998).
- Qin, S. J., and T. A. Badgwell, "An Overview of Industrial Model Predictive Control Technology," *Chemical Process Control*, Vol. V, J. C. Kantor, C. E. Garcia, and B. Carnahan, eds, CACHE, AIChE, New York, p. 232 (1997).
- Ramamurthi, Y., P. Sistu, and B. Bequette, "Control-Relevant Dynamic Data Reconciliation and Parameter Estimation," *Comput. Chem. Eng.*, **17**, 41 (1993).
- Rao, C. V., "Moving Horizon Strategies for the Constrained Monitoring and Control of Nonlinear Discrete-Time Systems," PhD Thesis, Univ. of Wisconsin-Madison (2000). <http://www.che.wisc.edu/~rao/group.ps>.
- Rao, C. V., and J. B. Rawlings, "Nonlinear Moving Horizon Estimation," *Int. Symp. on Nonlinear Model Predictive Control*, Ascona, Switzerland, (1998).
- Rao, C. V., J. B. Rawlings, and J. H. Lee, "Constrained Linear State Estimation—A Moving Horizon Approach," *Automatica*, **37**, 1619 (2001).
- Rao, C. V., J. B. Rawlings, and J. H. Lee, "Stability of Constrained Linear Moving Horizon Estimation," *Proc. American Control Conf.*, San Diego, CA (1999b).
- Rawlings, J. B., N. F. Jerome, J. W. Hamer, and T. M. Bruemmer, "Endpoint Control in Semi-Batch Chemical Reactors," *Proc. IFAC Symp. Dynamics and Control of Chemical Reactors, Distillation Columns, and Batch Processes*, p. 323 (1989).
- Robertson, D., "Development and Statistical Interpretation of Tools for Nonlinear Estimation," PhD Thesis, Auburn University, Auburn, AL (1996).
- Robertson, D. G., and J. H. Lee, "A Least Squares Formulation for State Estimation," *J. Process. Cont.*, **5**, 291 (1995).
- Robertson, D. G., and J. H. Lee, "On the Use of Constraints in Least Squares Estimation and Control," *Automatica*, in press (2002).
- Robertson, D. G., J. H. Lee, and J. B. Rawlings, "A Moving Horizon-Based Approach for Least-Squares State Estimation," *AIChE J.*, **42**, 2209 (1996).
- Russo, L. P., and R. E. Young, "Moving Horizon State Estimation Applied to an Industrial Polymerization Process," *Proc. 1999 American Control Conf.*, San Diego, CA (1999).
- Schuler, H., and C.-U. Schmidt, "Calorimetric-State Estimators for Chemical Reactor Diagnosis and Control: Review of Methods and Applications," *Chem. Eng. Sci.*, **47**, 899 (1992).
- Song, Y., and J. W. Grizzle, "The Extended Kalman Filter as a Local Asymptotic Observer for Discrete-Time Nonlinear Systems," *J. Math. Syst. Estim. and Control*, **5**, 59 (1995).
- Striebel, C., "Sufficient Statistics in the Optimum Control of Stochastic Systems," *J. Math. Anal. Appl.*, **12**, 576 (1965).
- Thomas, Y. A., "Linear Quadratic Optimal Estimation and Control with Receding Horizon," *Electron. Lett.*, **11**, 19 (1975).
- Tjoa, I. B., and L. T. Biegler, "Simultaneous Strategies for Data Reconciliation and Gross Error Detection of Nonlinear Systems," *Comput. Chem. Eng.*, **15**(10), 679 (1991).
- Tyler, M. L., "Performance Monitoring and Fault Detection in Control Systems," PhD Thesis, California Institute of Technology, Pasadena, CA (1997).
- Tyler, M. L., and M. Morari, "Stability of Constrained Moving Horizon Estimation Schemes," *AUT96-18* (Preprint), Automatic Control Laboratory, Swiss Federal Institute of Technology (1996).

Manuscript received Aug. 18, 2000, and revision received June 4, 2001.