A Moving Horizon-Based Approach for Least-Squares Estimation

Douglas G. Robertson and Jay H. Lee
Dept. of Chemical Engineering, Auburn University, Auburn, AL 36849

James B. Rawlings
Dept. of Chemical Engineering, The University of Wisconsin, Madison, WI 53706

A general formulation of the moving horizon estimator is presented. An algorithm with a fixed-size estimation window and constraints on states, disturbances, and measurement noise is developed, and a probabilistic interpretation is given. The moving horizon formulation requires only one more tuning parameter (horizon size) than many well-known approximate nonlinear filters such as extended Kalman filter (EFK), iterated EKF, Gaussian second-order filter, and statistically linearized filter. The choice of horizon size allows the user to achieve a compromise between the better performance of the batch least-squares solution and the reduced computational requirements of the approximate nonlinear filters. Specific issues relevant to linear and nonlinear systems are discussed with comparisons made to the Kalman filter, EKF, and other recursive and optimization-based estimation schemes.

Introduction

The estimation of model parameters and states is an integral part of many process modeling, monitoring, and control strategies. The success of many fault-detection methods and model-based controllers directly depends on the accuracy of the process model and the estimates of key states. The field of identification and estimation is usually divided into off-line and on-line methods. Off-line methods generally utilize batch processing of the data and are concerned with identification of time-invariant model parameters. The problem is often formulated as a least-squares optimization with the parameters chosen to minimize the sum of the squared errors between the measurements and model predictions (Hsia, 1977). On-line methods require recursive processing of the data and are concerned with time-varying parameters and states. For this problem, an accepted approach is a stochastic formulation in which parameters and states are modeled as random variables and the variance of the estimation error is minimized. This method is preferred over the deterministic least-squares formulation since it provides a framework in which the quality of estimation can be quantified.

The distinction between off-line (nonrecursive) and on-line (recursive) methods is not as rigid as it might seem. Most of the on-line methods were derived as recursive approximations to a particular off-line technique. Due to their computational advantages and ability to detect nonstationarities in the data and system properties, the use of recursive methods for off-line identification has been advocated by Ljung (1982). A proof of the convergence of recursive methods to the solution of the off-line problem for linear systems is given by Ljung (1982). For parameter estimation in state-space models (a nonlinear problem), convergence of the recursive prediction error method to the off-line solution is discussed in Ljung (1979).

For more general nonlinear dynamic models (see Eqs. 1 and 2), the distinction between recursive and nonrecursive methods becomes more blurred. Even for moderate-size models with moderate amounts of data, the off-line solution of the least-squares problem may not be computationally feasible. In this case, a recursive solution is the *only* alternative. Unfortunately, the exact recursive solution is infinite-dimensional (Kushner, 1964), and some approximations must be made. These approximations have led to various nonlinear filters (e.g., extended Kalman filter (EKF), iterated EKF, Gaussian second-order filter, and statistically linearized fil-

Correspondence concerning this article should be addressed to J. H. Lee.

ter); however, they often result in significant deviations from the off-line solution. This fact provides the motivation to examine a moving horizon formulation that allows the user to achieve a compromise between the approximate recursive filter and the batch least-squares solution at the cost of increased computational requirements.

Much of the current theory of stochastic estimation began with the seminal work of Kalman (Kalman, 1960; Kalman and Bucy, 1961), which provided a recursive solution to the minimum variance estimation problem for linear systems with Gaussian variables. The Kalman filter was extended to nonlinear systems (Kopp and Oxford, 1963; Cox, 1964) by linearizing the nonlinear model with a first-order Taylor expansion around the current estimate. However, when the error in the state estimates or the higher order terms neglected by the linear (first-order) model are significant, the EKF can exhibit poor convergence characteristics and biased estimates (Ljung, 1979; Maybeck, 1982).

As previously mentioned, least squares is recognized more as a technique for off-line identification, since its basic form is nonrecursive (which is easily understandable but unsuitable for real-time implementation). However, estimation of time-varying parameters and states can also be formulated as a least-squares problem, and recursive solutions can be developed for certain cases. For instance, several researchers (Albert and Sittler, 1966; Jazwinski, 1970) presented a recursive solution to the general linear state estimation problem using the least-squares approach and showed its equivalence to the Kalman filter under certain choices of the weighting matrices. This suggested a connection between the least-squares and stochastic formulations of state estimation. For example, using Bayes' Theorem, Cox (1964) gives the probabilistic interpretation that the least squares estimate corresponds to the maximum of the joint conditional density of the state trajectory.

While equivalent in its basic form to its stochastic counterpart (the Kalman filter), the least-squares formulation of state estimation enables us to address some additional issues with relative ease. For instance, when optimization software (quadratic programming or nonlinear programming) is used to solve the least-squares problem, inequality constraints can be placed on the unknown variables. This is useful from an engineering viewpoint since the prior knowledge of the process is often in the form of inequalities (e.g., variables such as temperature, pressure, flow rates, and concentrations, must be nonnegative and cannot go above some upper bound; the rate of change of these variables is also bounded by mass and energy balance considerations). It is difficult to use this information in the context of the Kalman filter since the filter calculates the state estimate from its probability distribution and does not explicitly compute the sequence of unknown variables (initial errors, disturbances, noise). To be interpreted as the minimum variance estimator, the Kalman filter implicitly assumes a normal distribution for the unknown random variables. By using constraints, these variables can be modeled as *truncated* normal variables. (The probability density function of a truncated normal random variable resembles the bell-shaped curve of its normal counterpart with the tails cut off.) Although not immediately obvious, the concept of a truncated normal variable offers a significant advantage in terms of the robustness of the estimates and modeling the random variables (we can model unimodal distributions that are very different from the normal distribution).

In addition, in the least-squares formulation, nonlinear model equations can be used explicitly—as opposed to the first-order approximations used by the EKF. However, in order to gain these additional benefits, it is necessary to minimize a least-squares objective subject to nonlinear ordinary differential equations (ODE) and other inequality constraints. Given the computational limits of the 1960s, it is not surprising that research in this area dwindled in comparison to recursive techniques based on linear filtering theory.

We believe that there are several factors that justify reexamination of this method as a tool for state/parameter estimation of nonlinear stochastic models. The steady advance in computer efficiency and numerical optimization methods coupled with decreases in hardware costs have contributed to the possibility of implementation of techniques that were once considered impractical. For example, successive quadratic programming (SQP) has greatly reduced the computational time for solving general nonlinear programming problems (Gill et al., 1981). Furthermore, by taking advantage of the fact that the first-order conditions for optimality for these types of problems is nearly block diagonal, Albuquerque and Biegler (1994) have proposed a method where the computational cost increases linearly with the size of the problem (as opposed to the quadratic to cubic rate of traditional SQP).

The popularity of optimization-based control methods like model predictive control (MPC) is also a strong influence. Given the duality between linear quadratic (LQ) regulator and the Kalman filter, least squares estimation provides a similar framework in which the Kalman filter can be tailored into a technique with flexibility and intuitive flavor. Another important feature in the analogy between least-squares estimation and MPC is that we can take advantage of the wealth of theoretical and practical knowledge gained through development of MPC algorithms (Muske and Rawlings, 1993a; Rawlings et al., 1994).

The problem addressed in this article is a general nonlinear estimation problem, which includes parameter estimation as a special case. Our objective is twofold. First, a general formulation of least-squares estimation along with its probabilistic interpretation is presented. The need for this becomes apparent when one considers the growing number of optimization-based estimation strategies proposed in recent years (Jang et al., 1986; Kim et al., 1991; Liebman et al., 1992; Ramamurthi et al., 1993). Understanding the probabilistic interpretation of the general formulation provides insight into the assumptions made by other researchers. Secondly, we present some new results within the suggested framework. Based on the probabilistic interpretation, a method employing a moving estimation window is proposed to control the size of the optimization. This is in the same spirit as the moving control horizon used by MPC and allows for a trade-off between the accuracy of the estimation and computational requirements. The incorporation of prior knowledge of the unknown variable ranges as inequality contraints is discussed and a probabilistic interpretation of the constrained formulation is given. Specific issues relevant to linear and nonlinear systems are discussed. Comparisons are made with the Kalman filter, extended Kalman filter, and other recursive and optimization-based estimation techniques.

Dynamic Model

Assume that the system is described by the following model:

$$\frac{dx_t}{dt} = f(x_t, p_t, u_t) + \omega_t \tag{1}$$

$$y_k = g(x_k, p_k) + \nu_k, \tag{2}$$

where x is the state vector, u a known input vector, p a vector representing unmeasured load disturbances and uncertain, (possibly) time-varying model parameters, ω a vector representing errors in the state equations, y the measurement vector, and ν a vector representing errors in the measurement equations. The subscript t denotes variables that are continuous in time, and the subscript k denotes values of these variables at discrete times. We assume that the sampling time of the process is 1, that is, y is observed and u is manipulated only at discrete time instants of $t = 0, 1, \ldots$.

To simplify the presentation, $\{\nu_k\}$ is modeled as a zero-mean, uncorrelated sequence of random vectors and p_k is modeled as a discrete random walk process:

$$p_k = p_{k-1} + w^p_{k-1}, \tag{3}$$

where $\{w^p_k\}$ is a zero-mean, uncorrelated sequence of random vectors. If an element of the parameter vector is assumed to be constant, then the corresponding element of w^p is identically zero. More complicated disturbance/noise models (to model more complex behavior for p, ω, and ν) can easily be incorporated into the formulation by augmenting the state/parameter equations.

The solution of Eq. 1 for one sampling interval is

$$x_k = x_{k-1} + \int_{k-1}^{k} f(x_t, p_{k-1}, u_{k-1})\,dt + \int_{k-1}^{k} \omega_t\,dt \tag{4}$$

and is denoted as

$$x_k = F(x_{k-1}, p_{k-1}, u_{k-1}) + w^x_{k-1}, \tag{5}$$

where $\{w^x_k\}$ is assumed to be a zero-mean, uncorrelated sequence of random vectors. Note that $F(\cdot)$ generally cannot be expressed in a closed form, but represents the solution of the nonlinear ODE over one sampling interval.

Equations 2, 3, and 5 are combined to represent the form of the model that is used in this article:

$$\begin{bmatrix} x_k \\ p_k \end{bmatrix} = \begin{bmatrix} F(x_{k-1}, p_{k-1}, u_{k-1}) \\ p_{k-1} \end{bmatrix} + \begin{bmatrix} w^x_{k-1} \\ w^p_{k-1} \end{bmatrix} \tag{6}$$

$$y_k = g(x_k, p_k) + \nu_k. \tag{7}$$

For the remainder of the article, X_k will denote the augmented state vector $[x_k^T\, p_k^T]^T$, and w_k will denote the augmented state noise vector $[(w^x_k)^T (w^p_k)^T]^T$.

Least-Squares Formulation

The objective of batch state estimation at time k can be stated as:

> Given an initial estimate $X_{1|0}(\triangleq [x_{1|0}^T\, p_{1|0}^T]^T)$, the measurement sequence $\{y_1, \ldots, y_k\}$, and the model, Eqs. 6–7, estimate the error in the initial estimate X^e_1 ($\triangleq X_1 - X_{1|0}$) and the unknown sequence $\{w_1, \ldots, w_{k-1}\}$.

(Throughout this article, we use the notation $\{\cdot\}_{\ell|k}$ to denote the estimate of a vector at time ℓ based on the measurements up to time k.) Once estimates of these unknown variables have been determined, the current state estimate is obtained via the model equations.

There are usually an infinite number of choices for $\{w_1, \ldots, w_{k-1}\}$, $\{\nu_1, \ldots, \nu_k\}$ and X^e_1 that are consistent with a given measurement sequence $\{y_1, \ldots, y_k\}$. (We note that given the measurement sequence, the random variables $\{\nu_1, \ldots, \nu_k\}$ are not independent of X^e_1 and $\{w_1, \ldots, w_{k-1}\}$.) It is therefore necessary to establish a criterion for calculating the "best" estimates of these unknown variables (or, equivalently, the states). If we consider the unknown random variables $\{w_1, \ldots, w_{k-1}\}$, $\{\nu_1, \ldots, \nu_k\}$ and X^e_1 to be errors in the state equations, measurement equations, and initial estimate, respectively, we can use classical least-squares theory to minimize these errors. In a sense, we are obtaining the "best fit" of the state trajectory to the observations subject to the model. The weighted least-squares estimate at time k can be obtained by minimizing the following quadratic function of the unknown variables:

$$\min_{\substack{X^e_1 \\ w_1, \ldots, w_{k-1}}} \quad (X^e_1)^T P^{-1}_{1|0} X^e_1 + \sum_{\ell=1}^{k} \nu_\ell^T R^{-1} \nu_\ell + \sum_{\ell=1}^{k-1} w_\ell^T Q^{-1} w_\ell \tag{8}$$

s.t.

$$\nu_\ell = y_\ell - g(x_\ell, p_\ell)$$

$$x_\ell = F(x_{\ell-1}, p_{\ell-1}, u_{\ell-1}) + w^x_{\ell-1}$$

$$p_\ell = p_{\ell-1} + w^p_{\ell-1},$$

where

$$X^e_1 \triangleq X_1 - X_{1|0}.$$

Recall that $x_\ell = F(x_{\ell-1}, p_{\ell-1}, u_{\ell-1}) + w^x_{\ell-1}$ is an ODE constraint. The positive definite weighting matrices R^{-1}, Q^{-1}, and $P^{-1}_{1|0}$ are quantitative measures of our confidence in the output model, the dynamical system model, and the initial estimate, respectively. Often, diagonal matrices are used with each diagonal element chosen to be inversely proportional to the expected squared magnitude of the corresponding unknown variable. If correlations among the variables are known to exist, they can be accounted for by using nondiagonal weighting matrices.

The solution of the preceding optimization represents the least-squares estimates for the unknowns and can be denoted as $X^e_{1|k}$, $w_{1|k}, \ldots, w_{k-1|k}$, and $\nu_{1|k}, \ldots, \nu_{k|k}$. The current state estimate $X_{k|k}$ can be computed by integrating the model equation, Eq. 6, with initial condition $X_{1|0} + X^e_{1|k}$ and input sequence $w_{1|k}, \ldots, w_{k-1|k}$.

Probabilistic interpretation

The preceding formulation provides little insight into choosing the weighting matrices or interpreting the quality of the estimates. By giving a statistical description to the unknown variables and choosing an appropriate criterion for determining what is meant by the "best" estimates, we arrive at a natural choice for the weighting matrices as well as a well-defined statistical interpretation of the state estimates. We have said that there are an infinite combination of the variables $\{X_1^e, w_1, \ldots, w_{k-1}, \nu_1, \ldots, \nu_k\}$ that explain a given measurement sequence. Therefore, we must provide additional information about the model prior to the estimation. The Bayesian framework provides a way of including this prior knowledge in a rigorous statistical manner. All of the information about the states that is contained in the measurement sequence can be written in terms of the conditional joint density function $p(X_1, \ldots, X_k \mid y_1, \ldots, y_k)$. While this density is generally complex and awkward to work with, Bayes' theorem provides a method of writing the conditional density in terms of the (often) less complicated densities of X_1^e, $\{w_k\}$, and $\{\nu_k\}$. When X_1^e, $\{w_k\}$, and $\{\nu_k\}$ are uncorrelated, zero-mean Gaussian sequences with covariance $P_{1|0}$, Q, and R, respectively, we can write the joint conditional probability density function of the state trajectory as

$$p(X_1, \ldots, X_k \mid y_1, \ldots, y_k) = c' \exp\left[-\frac{1}{2}\left((X_1^e)^T P_{1|0}^{-1} X_1^e + \sum_{\ell=1}^{k} \nu_\ell^T R^{-1} \nu_\ell + \sum_{\ell=1}^{k-1} w_\ell^T Q^{-1} w_\ell\right)\right], \quad (9)$$

where c' is a constant independent of $\{X_1, \ldots, X_k\}$ (for proof, see Cox, 1964). It is obvious that the maximum of the conditional probability density is obtained by minimizing the least-squares objective given by Eq. 8. Thus the least squares estimate corresponds to maximizing the joint conditional probability density function with respect to $\{X_1, \ldots, X_k\}$. The estimate is the peak or mode of the joint conditional density. Since the conditional density in Bayesian estimation is often called the posterior density, the estimate that maximizes this density is called the *maximum a posteriori* (MAP) estimate. Bayes' theorem also provides a means of handling non-Gaussian systems (see, for example, Ho and Lee, 1964; Friedland and Bernstein, 1966) with an associated increase in the complexity of the conditional density.

Moving horizon formulation

The size of the optimization in Eq. 8 increases linearly with the number of measurements. For an estimation technique to be computationally feasible, we must be able to bound the number of variables to be estimated. The batch estimation problem can be modified to employ a fixed-size moving window in which the number of measurements that we base our estimate on (and, hence, the size of the optimization) remains constant. The moving horizon state estimation problem at time k with horizon size of m (the horizon size is equal to the number of measurements used) is formulated as follows:

$$\min_{\substack{X_{k-m+1}^e \\ w_{k-m+1}, \ldots, w_{k-1}}} (X_{k-m+1}^e)^T P_{k-m+1|k-m}^{-1} X_{k-m+1}^e + \sum_{\ell=k-m+1}^{k} \nu_\ell^T R^{-1} \nu_\ell + \sum_{\ell=k-m+1}^{k-1} w_\ell^T Q^{-1} w_\ell \quad (10)$$

s.t.

$$\nu_\ell = y_\ell - g(x_\ell, p_\ell)$$
$$x_\ell = F(x_{\ell-1}, u_{\ell-1}, p_{\ell-1}) + w_{\ell-1}^x$$
$$p_\ell = p_{\ell-1} + w_{\ell-1},$$

where

$$X_{k-m+1}^e \triangleq X_{k-m+1} - X_{k-m+1|k-m},$$

where $X_{k-m+1|k-m}$ represents the least-squares estimate of X_{k-m+1} obtained at time $k-m$ and $P_{k-m+1|k-m}^{-1}$ is the weighting matrix expressing the confidence in the estimate (e.g., inverse of the conditional covariance of X_{k-m+1} at time $k-m$). At the beginning of the estimation, the number of measurements is allowed to grow until it reaches the size of the horizon (i.e., $t=m$). At the next time step the initial estimate $X_{1|0}$ is replaced by $X_{2|1}$ and the weighting matrix $P_{1|0}^{-1}$ is replaced by $P_{2|1}^{-1}$. The first measurement y_1 is discarded as the current measurement y_{m+1} is made available. This procedure is repeated at each time step, and the optimization remains at constant size for all future times. The probabilistic interpretation of $P_{k|k-1}$ as the covariance of $X_{k|k-1}$ provides us with a basis for updating $P_{2|1}$ from $P_{1|0}$.

Constrained least-squares estimation

The states, parameters, and to some extent $\{\nu_k\}$ and $\{w_k\}$ correspond to physical characteristics of the system. Although formulating a probabilistic model for these variables is difficult, an engineer will usually have knowledge about the range of values that they can assume. This information can be used as constraints in the least-squares objective to improve the estimation algorithm (Muske et al., 1993b). We may use some or all of the following constraint specifications:

$$\nu_{\min} \le \nu_\ell \le \nu_{\max} \quad \text{for } k-m+1 \le \ell \le k$$
$$x_{\min} \le x_\ell \le x_{\max} \quad \text{for } k-m+1 \le \ell \le k$$
$$p_{\min} \le p_\ell \le p_{\max} \quad \text{for } k-m+1 \le \ell \le k$$
$$w_{\min} \le w_\ell \le w_{\max} \quad \text{for } k-m+1 \le \ell \le k-1.$$

Of course, more complicated constraint formulations (including functions of one or more of the variables) are possible.

Probabilistic interpretation of constrained least-squares estimation

The Bayesian interpretation of the Kalman filter requires a normal distribution for the unknown random variables. This means that although the probability is small, these variables may assume arbitrarily large values. For real systems it is usually the case that the unknown variables have zero probability beyond certain ranges. In other words, variables representing physical quantities are almost always bounded. This means that the basic assumption of normally distributed random variables (or for that matter, any distribution with infinite support) is incorrect. While this is not a problem for the ideal case, model and covariance errors inherent in real systems can lead to the Kalman filter providing estimates that, in fact, have zero probability. A more realistic assumption for the distribution of the unknown random variables is the *truncated* normal distribution. A complete discussion of the advantages of this concept is deferred until a future article. At this point it is enough to analyze the effect of constraints on the statistical interpretation of the least-squares estimates.

The discussion of the probabilistic interpretation of constraints is divided into two parts: constraints on the unknown random variables and constraints on the states and parameters. Constraints on the unknown random variables are of the form

$$\begin{aligned} v_{\min} \le v_\ell \le v_{\max} &\quad \text{for } k-m+1 \le \ell \le k \\ w_{\min} \le w_\ell \le w_{\max} &\quad \text{for } k-m+1 \le \ell \le k-1 \\ X^e_{\min} \le X^e_{k-m+1} &\le X^e_{\max}. \end{aligned}$$

Note that (1) when the disturbance/parameter vector p_ℓ is modeled as integrated white noise, limits on Δp_ℓ correspond to limits on the magnitude of w_ℓ^p; and (2) constraints on the initial state estimate correspond to constraints on X^e_{k-m+1}. The solution of Eq. 10 subject to the preceding constraints *preserves* the MAP interpretation of the state estimates. This can be shown by deriving the conditional joint probability density function and observing that it is in the same form as Eq. 9. In this case c' is a positive constant independent of $\{X_{k-m+1}, \dots, X_k\}$ whenever the unknown random variables are within the given bounds and is zero otherwise (Proof: See Appendix).

When constraints are placed on the states and parameters:

$$\begin{aligned} x_{\min} \le x_\ell \le x_{\max} &\quad \text{for } k-m+1 \le \ell \le k \\ p_{\min} \le p_\ell \le p_{\max} &\quad \text{for } k-m+1 \le \ell \le k, \end{aligned}$$

the unknown random variables $\{X^e_{k-m+1}, w_{k-m+1}, \dots, w_{k-1}\}$ may no longer be independent. For example, if the estimate of p_ℓ $(k-m+1 < \ell < k)$ is at its upper limit, then from Eq. 3 w_ℓ^p cannot be positive. From Eq. 10, p_ℓ explicitly depends on $X^e_{k-m+1}, w_{k-M+1}, w_{k-m+2}, w_{\ell-1}$; therefore, these variables are no longer independent. Whether or not the least-squares solution is still the MAP estimate depends on how we choose to model this dependence (it is more a question of modeling philosophy than statistics) and is discussed in the Appendix.

Linear State Estimation

When the model given by Eqs. 1–2 is linear, the application of the method and the analysis of the results are simplified. Since the model constraint is linear, the explicit solution for the least-squares estimate can be derived. When inequality constraints are used to bound the estimates, a quadratic program can be used to obtain the solution. Assume that the linear model has the following form:

$$X_k = \Phi X_{k-1} + \Gamma u_{k-1} + w_{k-1} \tag{11}$$

$$y_k = \Xi X_k + v_k. \tag{12}$$

The moving horizon estimate for the linear model minimizes Eq. 10 subject to the preceding linear model constraints.

An important issue that has not yet been addressed is how to update the initial estimate $X_{k-m+1|k-m}$ and its weighting matrix $P^{-1}_{k-m+1|k-m}$. Based on a past horizon of m measurements, $X_{k-m+1|k-m}$ is the value of X_{k-m+1} that jointly maximizes

$$p(X_{k-2m+1}, \dots, X_{k-m+1} | y_{k-2m+1}, \dots, y_{k-m}).$$

From the discussion on the Kalman filter below (for the unconstrained case),

$$\begin{aligned} X_{k-m+1|k-m} &= \mathrm{E}[X_{k-m+1} | X_{k-m|k-m-1}, P_{k-m|k-m-1}, y_{k-m}] \\ &= \mathrm{E}[X_{k-m+1} | X_{k-m|k-m}, P_{k-m|k-m}] \\ &= \Phi X_{k-m|k-m} + \Gamma u_{k-m} \end{aligned} \tag{13}$$

where $X_{k-m|k-m}$ is obtained from the solution of the least-squares problem at time $k-m$, and E denotes expectation. Given the probabilistic interpretation of $P_{k-m+1|k-m}$ as the covariance of $(X_{k-m+1} - X_{k-m+1|k-m})$, we can calculate $P_{k-m+1|k-m}$ from $P_{k-m|k-m-1}$ using linear filtering theory:

$$P_{k-m|k-m} = P_{k-m|k-m-1} - P_{k-m|k-m-1}\Xi^T \times (\Xi P_{k-m|k-m-1}\Xi^T + R)^{-1} \Xi P_{k-m|k-m-1} \tag{14}$$

$$P_{k-m+1|k-m} = \Phi P_{k-m|k-m}\Phi^T + Q. \tag{15}$$

The matrix inversion lemma can be used to rewrite Eqs. 14 and 15 in terms of the inverse matrices used in the least-squares objective:

$$P^{-1}_{k-m|k-m} = P^{-1}_{k-m|k-m-1} + \Xi^T R^{-1} \Xi \tag{16}$$

$$P^{-1}_{k-m+1|k-m} = Q^{-1} - Q^{-1}\Phi(P^{-1}_{k-m|k-m} + \Phi^T Q^{-1}\Phi)^{-1}\Phi^T Q^{-1}. \tag{17}$$

Equations 13, 16 and 17 are then used to update the initial estimate and weighting matrix at each time step.

The use of inequality constraints to bound the unknown variables implies that they are not normally distributed. When the constraints lead to distributions of the unknown variables that are significantly non-Gaussian (e.g., constraints within three standard deviations of the mean), the update equations just given—which are based on the normal distribution—are no longer optimal. In this case, the update of the initial estimate is more complicated. As in the case of nonlinear estimation discussed below, since the conditional distribution of the state is no longer Gaussian, some approximations must be made. In addition, the constraints will define a feasible region for the state estimates that must be propagated using the model equations. Discussion of these issues is beyond the scope of this article and will be treated in a future article.

Nonlinear Estimation

When a linear model is not adequate to describe the behavior of the system, nonlinear estimation is required. In addition, even for linear models, simultaneous state/parameter estimation (required in the case of unknown or time-varying parameters) leads to a nonlinear estimation problem. The moving horizon state estimation problem at time k for a horizon size of m is given in Eq. 10. The solution requires an ODE solver coupled with a nonlinear optimization algorithm and is computationally demanding. A better method is to discretize the ODE constraints using weighted residual methods (e.g., orthogonal collocation, see Villadsen and Michelson, 1978) and express them as a set of algebraic constraints. The optimization is then a nonlinear program (NLP) and can be solved via techniques such as SQP. The nonlinear state/parameter estimation technique is dual to the nonlinear MPC techniques previously proposed (Cuthrell and Biegler, 1987; Rawlings et al., 1994). Hence, various discretization strategies and solution procedures used in nonlinear MPC are directly applicable to estimation as well.

For nonlinear systems, updating of the weighting matrix, $P_{k-m+1|k-m}^{-1}$, is more complicated than the linear case. To maintain the *maximum a-posteriori* interpretation of our estimates, we assume that the initial state is normally distributed with mean $X_{1|0}$ and variance $P_{1|0}$. When the number of measurements equals the horizon size, we must update $X_{1|0}$ and $P_{1|0}^{-1}$ with $X_{2|1}$ and $P_{2|1}^{-1}$ at the next time step. However, due to the nonlinearity of the system, the error in $X_{2|1}$ is no longer normally distributed. This means that the distribution of X_2 can no longer be completely characterized by its mean and covariance (or, in general, any other finite set of parameters; Jazwinski, 1970). The exact solution for the evolution of the conditional density (Kushner, 1964) requires an infinite-dimensional system of equations. Therefore, to obtain a computationally feasible algorithm, some approximations must be made. A great deal of research in nonlinear estimation has focused on this subject in the context of recursive estimation. One approach is to expand nonlinear equations in a Taylor series around the conditional mean to obtain equations for the evolution of the mean and covariance. When only first-order terms are retained the procedure is called the extended Kalman filter. Several different second-order filters are discussed in Jazwinski (1970). The statistically linearized filter (Gelb, 1974) accounts for the uncertainty in the estimates in obtaining a first-order approximation of the nonlinear equations and is generally more accurate than the Taylor series expansion. Any of the methods just cited could be used for the moving horizon estimator since they are all based on propagating the mean and covariance of the conditional density. This issue is discussed below in the sections on the extended Kalman filter and other approximate nonlinear filters.

Comparison with Other Estimation Methods

In this section we show how other methods of state estimation can be represented in the framework of least squares estimation. The Kalman filter and its equivalence to least squares is widely known, and there is no need to rederive this result. The purpose of the first subsection is to highlight some aspects of the linear filtering problem which contrast the approximate nonlinear filters of the following sections. In the next subsection the equivalence between the moving horizon estimator and EKF is discussed. The following subsection discusses the equivalence between the moving horizon estimator and other approximate nonlinear filters. The final subsection presents a comparison between the moving horizon estimator discussed in this article and previous optimization-based approaches.

Kalman filter

Instead of finding the state estimates that maximize the conditional density, another approach is to choose the estimate with the minimum error variance. In this case the loss function becomes

$$\mathrm{E}\left[(X_{k+1}-X_{k+1|k})^T(X_{k+1}-X_{k+1|k})\right], \tag{18}$$

where E denotes expectation, X_{k+1} is the true state at time $k+1$, and $X_{k+1|k}$ is the state estimate based on the measurements up to time k. (Note that we are using the "predicted" estimate $X_{k+1|k}$ instead of the "filtered" estimated $X_{k+1|k+1}$. The reason for this is that it simplifies comparisons between the moving horizon estimator and the EKF and other approximate nonlinear filters in the next two subsections.) The optimal solution is the minimum variance estimate (which is an attractive property for the estimates to possess). It is well known that the minimum variance estimate is simply the conditional mean:

$$X_{k+1|k}=\mathrm{E}[X_{k+1}|y_1, \ldots, y_k]. \tag{19}$$

Another important property of an estimation algorithm is its ability to be implemented in a recursive form. As the number of measurements increases, the expression for the conditional expectation in Eq. 19 becomes increasingly complex. Since the process is Markov (the future states depend only on the present and not the past states), the conditional probability density $p(X_k|y_1, \ldots, y_{k-1})$ represents all of the information contained is the measurement sequence $\{y_1, \ldots, y_{k-1}\}$. We can therefore rewrite Eq. 19 as

$$X_{k+1|k}=\mathrm{E}[X_{k+1}|p(X_k|y_1, \ldots, y_{k-1}), y_k]. \tag{20}$$

Depending on the complexity of $p(X_k|y_1, \ldots, y_{k-1})$, this can represent a significant reduction in the amount of data required to compute the current estimate. Suppose that $p\,(X_k|y_1, \ldots, y_{k-1})$ is normal. Three important facts about normal random variables that simplify this objective are (1) the probability distribution of a normal random variable is completely determined by its mean and covariance; (2) linear combinations of normal random variables have a normal distribution; and (3) if two random variables z_1 and z_2, say, have a jointly normal distribution, then the conditional distribution $p(z_1 \mid z_2)$ is normal. The first fact states that $p(X_k \mid y_1, \ldots, y_{k-1})$ can be completely characterized by its mean $X_{k \mid k-1}$ and covariance $P_{k \mid k-1} = \mathrm{E}[(X_k - X_{k \mid k-1})(X_k - X_{k \mid k-1})^T]$. The second fact implies that for *linear systems* the joint distribution $p[X_{k+1}, y_k \mid p(X_k \mid y_1, \ldots, y_{k-1})]$ is normal. The third fact states that $p\,[X_{k+1} \mid p(X_k \mid y_1, \ldots, y_{k-1}), y_k]$ is normal. Based on these facts we can rewrite Eq. 20 as

$$X_{k+1 \mid k} = \mathrm{E}[X_{k+1} \mid X_{k \mid k-1}, P_{k \mid k-1}, y_k]. \tag{21}$$

Since $X_{1 \mid 0}$ is normally distributed, $X_{k+1 \mid k}$ is normal by induction. In other words, for linear Gaussian systems the statistics $X_{k \mid k-1}$ and $P_{k \mid k-1}$ summarize all of the information contained in the measurement sequence $\{y_1, \ldots, y_{k-1}\}$. The recursive solution to the minimum variance estimation problem was given by Kalman (1960, 1961) under the assumption that the unknown random variables are normally distributed.

We have stated that the state estimate obtained from the Kalman filter corresponds to the conditional mean. Recall that the state estimate obtained by least-squares estimation corresponds to the mode of the conditional joint density

$$p(X_1, \ldots, X_k \mid y_1, \ldots, y_k). \tag{22}$$

Since the marginal distribution of jointly normal random variables is also normal (with the same mode), the value of X_k that corresponds to the mode of Eq. 22 will also be the mode of

$$p(X_{k-m+1}, \ldots, X_k \mid y_1, \ldots, y_k), \tag{23}$$

which, as discussed earlier, for linear, Gaussian systems is equivalent to

$$p(X_{k-m+1}, \ldots, X_k \mid X_{k-m+1 \mid k-m}, P_{k-m+1 \mid k-m}, y_{k-m+1}, \ldots, y_k), \tag{24}$$

which is the conditional density maximized by the solution of the moving horizon objective function. Since m is arbitrary, obviously the current estimate $X_{k+1 \mid k}$ does not depend on the horizon size. For linear systems, the conditional density is the familiar bell-shaped normal density. Since this density is symmetric and unimodal, the mean also corresponds to the mode; therefore, for linear systems the least-squares estimator and the Kalman filter are equivalent (regardless of horizon size). For linear systems, the only advantage of least-squares estimation (besides its intuitive appeal) is its ability to incorporate constraints as previously discussed.

Extended Kalman filter

Unfortunately, nonlinear combinations of normal random variables do not have a normal distribution. We have mentioned several approximate nonlinear filters that are compatible with the moving horizon estimator. The EKF is by far the most popular and has been used in the past for comparison with other horizon-based estimators. For these reasons, the propagation of the initial estimate and weighting matrix for the moving horizon estimator used in this article (e.g., for simulation) is based on the EKF algorithm. This subsection provides a detailed discussion of how the equivalence between the EKF and moving horizon estimator is established. The following subsection discusses how these ideas are extended to other approximate nonlinear algorithms.

The EKF can be divided into two steps. The first is termed the measurement correction (given $X_{k \mid k-1}$, its error covariance ($P_{k \mid k-1}$), and the current measurement y_k, find $X_{k \mid k}$ and $P_{k \mid k}$) and the second is the model prediction (given $X_{k \mid k}$ and $P_{k \mid k}$, find $X_{k+1 \mid k}$ and $P_{k+1 \mid k}$). The steps are then repeated for all future times.

Measurement Correction.. Recall that we are given $X_{k \mid k-1}$, $P_{k \mid k-1}$, and y_k. Define as before

$$X_k^e \triangleq X_k - X_{k \mid k-1}. \tag{25}$$

If X_k^e is small, then a Taylor expansion of the measurement equation, Eq. 7, around $X_{k \mid k-1}$ yields

$$y_k \approx g(X_{k \mid k-1}) + \Xi_{k \mid k-1} X_k^e + \nu_k, \tag{26}$$

where

$$\Xi_{k \mid k-1} = \left.\frac{\partial g(X)}{\partial X}\right|_{X = X_{k \mid k-1}}. \tag{26}$$

The approximate measurement equation is linear in the unknown variable X_k^e. Furthermore, from Eqs. 21 and 25,

$$\mathrm{E}[X_k^e \mid X_{k \mid k-1}, P_{k \mid k-1}] = 0 \tag{27}$$

$$\mathrm{cov}(X_k^e) = P_{k \mid k-1}. \tag{28}$$

Since X_k^e and y_k are jointly normal, we can use the Kalman filter to calculate $\mathrm{E}[X_k^e \mid X_{k \mid k-1}, P_{k \mid k-1}, y_k]$ and its covariance $P_{k \mid k}$. (Note that the EKF uses the linear Kalman filter to estimate the *deviation* from a reference state.) Equation 25 is used to obtain the measurement correction equations:

$$X_{k \mid k} = X_{k \mid k-1} + L_k[y_k - g(X_{k \mid k-1})] \tag{29}$$

$$P_{k \mid k} = (I - L_k \Xi_{k \mid k-1}) P_{k \mid k-1}, \tag{30}$$

where

$$\begin{aligned} L_k &= P_{k \mid k-1} \Xi_{k \mid k-1}^T (\Xi_{k \mid k-1} P_{k \mid k-1} \Xi_{k \mid k-1}^T + R)^{-1} \\ R &= \mathrm{cov}\,(\nu_k) \end{aligned} \tag{31}$$

Model Prediction. Proceeding as before, define

$$\tilde{X}_k^e \triangleq X_k - X_{k \mid k}. \tag{32}$$

If $\tilde{X}_k^e$ is small, then a Taylor expansion of the model equation, Eq. 6, around $X_{k \mid k}$ yields

$$\begin{bmatrix} x_{k+1} \\ p_{k+1} \end{bmatrix} \approx \begin{bmatrix} F(x_{k \mid k}, p_{k \mid k}, u_k) \\ p_{k \mid k} \end{bmatrix} + \Phi_{k \mid k} \begin{bmatrix} \tilde{x}_k^e \\ \tilde{p}_k^e \end{bmatrix} + \begin{bmatrix} w_{k-1}^x \\ w_{k-1}^p \end{bmatrix}, \tag{33}$$

where

$$\Phi_{k|k} = \begin{bmatrix} A_{k|k} & B^p_{k|k} \\ 0 & I \end{bmatrix}$$

$$A_{k|k} = \exp(\tilde{A}_{k|k} \cdot T_s)$$

$$\tilde{A}_{k|k} = \left. \frac{\partial f(x,p,u)}{\partial x} \right|_{x = x_{k|k},\, p = p_{k|k},\, u = u_k}$$

$$B^p_{k|k} = \int_0^{T_s} \exp(\tilde{A}_{k|k} \cdot \tau)\, d\tau \cdot \tilde{B}^p_{k|k}$$

$$\tilde{B}^p_{k|k} = \left. \frac{\partial f(x,p,u)}{\partial p} \right|_{x = x_{k|k},\, p = p_{k|k},\, u = u_k},$$

where T_s is the sampling time. The approximate state equation is linear in the unknown variable $\tilde{X}^e_k$. Once again,

$$E[\tilde{X}^e_k \,|\, X_{k|k}, P_{k|k}] = 0 \tag{34}$$

$$\operatorname{cov}(\tilde{X}^e_k) = P_{k|k}, \tag{35}$$

and we obtain the model update equations:

$$\begin{bmatrix} x_{k+1|k} \\ p_{k+1|k} \end{bmatrix} = \begin{bmatrix} F(x_{k|k}, p_{k|k}, u_k) \\ p_{k|k} \end{bmatrix} \tag{36}$$

$$P_{k+1|k} = \Phi_{k|k} P_{k|k} \Phi^T_{k|k} + Q, \tag{37}$$

where

$$Q = \operatorname{cov}(w_k).$$

Equations 29–31 and 36–37 compose the extended Kalman filter. The derivation required the assumption that X^e_k and $\tilde{X}^e_k$ are small (or, equivalently, that the estimates are close to the true values). This will not be true, in general, if $P_{k|k}$ or Q is large (Jazwinski, 1970). Even if the estimation error is small, we have no guarantee that the linearized model is a good approximation of the nonlinear system.

The same concepts used in deriving the extended Kalman filter can be applied to the initial state $X_{k-m+1|k-m}$ and weighting matrix $P_{k-m+1|k-m}$ of the moving horizon estimator to calculate $X_{k-m+2|k-m+1}$ and $P_{k-m+2|k-m+1}$ at the next time step—the major difference being that for the moving horizon estimator this update is not required until time k. When the the horizon size m is greater than one, we can take advantage of this fact by taking the Taylor expansion with respect to the *smoothed* estimate, $X_{k-m+1|k}$. This is done as follows, beginning with the measurement correction:

$$y_{k-m+1} \approx g(X_{k-m+1|k}) + \Xi_{k-m+1|k}(X_{k-m+1} - X_{k-m+1|k}) + \nu_{k-m+1}, \tag{38}$$

where

$$\Xi_{k-m+1|k} = \left. \frac{\partial g(X)}{\partial X} \right|_{X = X_{k-m+1|k}}.$$

An important fact to notice is that the covariance of the term $(X_{k-m+1} - X_{k-m+1|k})$ is $P_{k-m+1|k}$; whereas, it is necessary to express the error in terms of $(X_{k-m+1} - X_{k-m+1|k-1})$, which has covariance $P_{k-m+1|k-m}$. [To understand why this must be so, recall from the section on the Kalman filter that with a change in the time subscripts, $X_{k-m+1|k-m}$ and $P_{k-m+1|k-m}$ represent the information about X_{k-m+1} contained in $\{y_1, \ldots, y_{k-m}\}$. Since the measurements $\{y_{k-m+1}, \ldots, y_k\}$ appear explicitly in the moving horizon formulation at time k, the initial estimate of X_{k-m+1} should not be based on information contained in these measurements (otherwise, we would be using these measurements twice).] Equation 38 can be rewritten as

$$\begin{aligned} y_{k-m+1} &\approx g(X_{k-m+1|k})\Xi_{k-m+1|k}(X_{k-m+1} - X_{k-m+1|k} \\ &\qquad + X_{k-m+1|k-m} - X_{k-m+1|k-m}) + \nu_{k-m+1} \\ &= g(X_{k-m+1|k}) + \Xi_{k-m+1|k}(X_{k-m+1|k-m} - X_{k-m+1|k}) \\ &\qquad \Xi_{k-m+1|k}(X_{k-m+1} - X_{k-m+1|k-m}) \\ &\qquad + \Xi_{k-m+1|k-1}(X_{k-m+1} - X_{k-m+1|k-m}) + \nu_{k-m+1} \\ &= g(X_{k-m+1|k}) + \Xi_{k-m+1|k}(X_{k-m+1|k-m} - X_{k-m+1|k}) \\ &\qquad + \Xi_{k-m+1|k} X^e_{k-m+1} + \nu_{k-m+1}, \end{aligned} \tag{39}$$

where the first two terms in Eq. 39 are the predicted measurement and X^e_{k-m+1} is defined exactly as in the measurement-correction step of the extended Kalman filter. The measurement-correction terms for the (EKF-based) moving horizon estimator are

$$\begin{aligned} X_{k-m+1|k-m+1} &= X_{k-m+1|k-m} \\ &\quad + L_{k-m+1}[y_{k-m+1} - g(X_{k-m+1|k}) \\ &\quad - \Xi_{k-m+1|k}(X_{k-m+1|k-m} - X_{k-m+1|k})] \end{aligned} \tag{40}$$

$$P_{k-m+1|k-m+1} = (I - L_{k-m+1}\Xi_{k-m+1|k})P_{k-m+1|k-m} \tag{41}$$

$$\begin{aligned} L_{k-m+1} &= P_{k-m+1|k-m}\Xi^T_{k-m+1|k} \\ &\quad (\Xi_{k-m+1|k}P_{k-m+1|k-m}\Xi^T_{k-m+1|k} + R)^{-1}. \end{aligned} \tag{42}$$

Likewise, the Taylor expansion of the model equations around the smoothed estimate yields

$$\begin{aligned} \begin{bmatrix} x_{k-m+2} \\ p_{k-m+2} \end{bmatrix} &\approx \begin{bmatrix} F(x_{k-m+1|k}, p_{k-m+1|k}, u_{k-m+1}) \\ p_{k-m+1|k} \end{bmatrix} \\ &\quad + \Phi_{k-m+1|k} \begin{bmatrix} x_{k-m+1} - x_{k-m+1|k} \\ p_{k-m+1} - p_{k-m+1|k} \end{bmatrix} + \begin{bmatrix} w^x_{k-m+1} \\ w^p_{k-m+1} \end{bmatrix} \\ &= \begin{bmatrix} F(x_{k-m+1|k}, p_{k-m+1|k}, u_{k-m+1}) \\ p_{k-m+1|k} \end{bmatrix} \\ &\quad + \Phi_{k-m+1|k} \begin{bmatrix} x_{k-m+1|k-m+1} - x_{k-m+1|k} \\ p_{k-m+1|k-m+1} - p_{k-m+1|k} \end{bmatrix} \\ &\quad + \Phi_{k-m+1|k}\tilde{X}^e_{k-m+1} + \begin{bmatrix} w^x_{k-m+1} \\ w^p_{k-m+1} \end{bmatrix}, \end{aligned} \tag{43}$$

where

$$\Phi_{k-m+1|k} = \begin{bmatrix} A_{k-m+1|k} & B^p_{k-m+1|k} \\ 0 & I \end{bmatrix}$$

$$A_{k-m+1|k} = \exp(\tilde{A}_{k-m+1|k} \cdot T_s);$$

$$\tilde{A}_{k-m+1|k} = \left.\frac{\partial f(x,p,u)}{\partial x}\right|_{x=x_{k-m+1|k},\, p=p_{k-m+1|k},\, u=u_{k-m+1}}$$

$$B^p_{k-m+1|k} = \int_0^{T_s} \exp(\tilde{A}_{k-m+1|k}\cdot\tau)\, d\tau \cdot \tilde{B}^p_{k-m+1|k}$$

$$\tilde{B}^p_{k-m+1|k} = \left.\frac{\partial f(x,p,u)}{\partial p}\right|_{x=x_{k-m+1|k},\, p=p_{k-m+1|k},\, u=u_{k-m+1}}$$

$$\tilde{X}^e_{k-m+1} = \begin{bmatrix} x_{k-m+1} - x_{k-m+1|k-m+1} \\ p_{k-m+1} - p_{k-m+1|k-m+1} \end{bmatrix}.$$

The first two terms of Eq. 43 are the model prediction, and $\tilde{X}^e_{k-m+1}$ in the third term is defined exactly as in the measurement-prediction step of the extended Kalman filter. The measurement-prediction terms for the (EKF-based) moving horizon estimator are

$$\begin{bmatrix} x_{k-m+2|k-m+1} \\ p_{k-m+2|k-m+1} \end{bmatrix} = \begin{bmatrix} F(x_{k-m+1|k}, p_{k-m+1|k}, u_{k-m+1}) \\ p_{k-m+1|k} \end{bmatrix} + \Phi_{k-m+1|k} \begin{bmatrix} x_{k-m+1|k-m+1} - x_{k-m+1|k} \\ p_{k-m+1|k-m+1} - p_{k-m+1|k} \end{bmatrix} \tag{44}$$

$$P_{k-m+2|k-m+1} = \Phi_{k-m+1|k} P_{k-m+1|k-m+1} \Phi^T_{k-m+1|k} + Q. \tag{45}$$

Equations 40–42 and 44–45 compose the update equations for the EKF-based moving horizon estimator. When these equations are used for the update, the moving horizon estimator with horizon size of one ($m=1$) given by the least-squares objective of Eq. 10 is *equivalent* to the EKF when the measurement equation is linear (as is often the case). When the measurement equation is nonlinear, equivalence is maintained when the Taylor expansion of Eq. 26 is used instead. To show the equivalence we rewrite Eq. 10 for a linearized measurement equation and a horizon of one.

$$\min_{X^e_k} \; (X^e_k)^T P^{-1}_{k|k-1} X^e_k + \nu^T_k R^{-1} \nu_k \tag{46}$$

s.t.

$$\nu_k = y_k - g(X_{k|k-1}) - \Xi_{k|k-1} X^e_k,$$

where

$$\Xi_{k|k-1} = \left.\frac{\partial g(X)}{\partial X}\right|_{X=X_{k|k-1}}$$

$$X^e_k \overset{\Delta}{=} X_k - X_{k|k-1}. \tag{47}$$

Note that (1) the state equations do not appear in the formulation when $m=1$; (2) the problem is a *linear* least-squares problem; and (3) the solution corresponds to the measurement-correction step of the EKF. We are given $X_{k|k-1}$, $\mathrm{P}_{k|k-1}$, and y_k. The solution of the preceding least-squares problem gives $X^e_{k|k}$, which we use in Eq. 47 to calculate $X_{k|k}$. It can be easily verified that $X^e_{k|k}$ is exactly the measurement correction term $L_k(y_k - g(X_{k|k-1}))$ appearing in Eq. 29 of the EKF. $P_{k|k}$ can be calculated from $P_{k|k-1}$ using Eqs. 41 and 42, which are equivalent to Eqs. 30 and 31 of the EKF when the horizon size m equals one. We have thus calculated $X_{k|k}$ and $P_{k|k}$. At the next time step $(k+1)$, the initial conditions for the least-squares estimator are $X_{k+1|k}$ with weighting matrix $P_{k+1|k}$. These values are calculated from $X_{k|k}$ and $P_{k|k}$ using. Eqs. 44 and 45, which are equivalent to the corresponding equations of the extended Kalman filter, Eqs. 36 and 37, when $m=1$.

We have shown that the EKF-based moving horizon estimator is equivalent to the EKF when the horizon size is one. Furthermore, the horizon size is the only additional tuning parameter required by the moving horizon estimator. We are now in a position to examine the effect of *increasing the horizon size:*

1. From the discussion in the section on the Kalman filter (and the analogous structure of the Kalman filter and EKF), at time k the information contained in the measurement sequence $\{y_1, \ldots, y_k\}$ is condensed by the EKF into the three statistics $X_{k|k-1}$, $P_{k|k-1}$, and y_k. Since the system is nonlinear, the conditional density is non-Gaussian and some information will be lost (depending on the degree of nonlinearity and the magnitude of the estimation errors, this loss can be severe). On the other hand, at time k the moving horizon estimator represents the information contained in the measurement sequence $\{y_1, \ldots, y_k\}$ by the $m+2$ statistics $X_{k-m+1|k-m}$, $P_{k-m+1|k-m}$, $y_{k-m+1}, \ldots, y_k$. *None* of the information contained in the last m measurements is lost, and the estimation is based on the nonlinear model over this measurement horizon.
2. Since the update of the initial estimate and weighting matrix is based on smoothed estimates (meaning the Taylor expansion is a better approximation), the initial state and weighting matrix of the moving horizon estimator are a more accurate description of the past information than those of the EKF.

The moving horizon estimator retains all of the most recent information and is more efficient at summarizing past information.

Other approximate nonlinear filters

One method for improving the performance of the EKF is, at each time step, to iteratively relinearize the model equations about the filtered estimate. For example, suppose the measurement equation is nonlinear. At time k the measurement correction step, Eq. 29, gives the estimate $X_{k|k}$. This estimate could be used to relinearize the measurement equation following the procedure outlined in Eqs. 38 and 39 for $m=1$. Equation 40 can then be used to calculate an improved estimate of $X_{k|k}$. This procedure can be repeated until the difference in successive estimates of $X_{k|k}$ is less than some convergence criterion. This method is called the iterated extended Kalman filter and is given in Theorem 8.2 of Jazwinski (1970). Section 9.7 of Jazwinski proves that this filter converges to the mode of the conditional density $p(X_k \mid X_{k|k-1}, P_{k|k-1}, y_k)$. Therefore, for a horizon size of one *and* nonlinear measurement equations, the EKF-based moving horizon estimator is *equivalent* to the iterated extended Kalman filter.

For nonlinear state equations a second possibility would be to use $X_{k|k}$ and linear smoothing theory to obtain the smoothed estimate $X_{k-1|k}$. The smoothed estimate could be used to relinearize the state equations following the procedure of Eq. 43 with $m = 2$. Equations 44 and 45 can then be used to obtain an improved estimate $X_{k|k-1}$ and covariance $P_{k|k-1}$. The measurement equations can then be relinearized around the filtered estimate $X_{k|k}$ using Eq. 39 with $m = 1$. Equations 40 and 42 can then be used to calculate an improved estimate of $X_{k|k}$. Again, this procedure can be repeated until some convergence criterion is satisfied. This method is called the iterated linear filter-smoother and is given in Theorem 8.3 of Jazwinski (1970). Section 9.7 of Jazwinski states that this filter converges to the mode of the conditional density $p(X_{k-1}, X_k | X_{k-1|k-1}, P_{k-1|k-1}, y_k)$. This is somewhat different than the moving horizon estimate which, for $m = 2$, is the mode of $p(X_{k-1}, X_k | X_{k-1|k-2}, P_{k-1|k-2}, y_{k-1}, y_k)$, although the iterated linear filter-smoother has been found through simulation to provide (in many cases) a good approximation to the moving horizon estimator with horizon size 2. We could, of course, take the iterated linear filter-smoother one step further and include y_{k-1} by relinearizing the measurement equation at time $k-1$ around the smoothed estimate $X_{k-1|k}$ with initial conditions $X_{k-1|k-2}$ and $P_{k-1|k-2}$ and then moving forward in time using Eqs. 40–42 and 44–45. We have also tried iterative filters that smooth (by linearizing around the best estimates of the state trajectory) m time steps back; however, we have observed from simulations that the errors due to the linear smoothing often tend to reduce the effectiveness of such an approach for improving the estimates of the current state. However, there are instances (especially in batch estimation) when this type of smoothing is useful for improving open-loop estimates of unobservable states through better estimates of the initial conditions (see Kozub and MacGregor, 1992, for another approach).

For parameter estimation in linear systems, Ljung (1979) analyzed the convergence properties of the EKF in terms of an associated differential equation and proposed a recursive prediction error method (RPEM) that has global convergence properties. For more general nonlinear estimation problems, the significance of bias at each step of the EKF is often assessed by comparing the EKF to higher order filters such as those based on a second-order Taylor expansion—the terms neglected by the EKF representing the bias. Maybeck (1982) and Jazwinski (1970) discuss second-order filters and compare them to the EKF. It should be obvious now that a second-order Taylor expansion can be used to propagate the initial state $X_{k-m+1|k-m}$ and weighting matrix $P_{k-m+1|k-m}$ of the moving horizon estimator, resulting in a least-squares filter that is equivalent to the corresponding second-order filter when $m = 1$.

In general, a moving horizon estimator based on any approximate nonlinear filter that propagates the conditional mean and covariance can be constructed by following the procedure of the previous subsection (i.e., making all approximations with respect to the smoothed estimates $X_{k-m+1|k}$ and then manipulating the results to keep the errors in terms of the predicted estimates $X_{k-m+1|k-m}$). The parameter m represents the trade-off between additional computational requirements and improved estimation.

The effects of nonlinearities on the solution of the batch least-squares problem are also well known. The literature on nonlinear regression provides insight into the differences between first-order and second-order approximations. Although these results were derived for the better-known nonlinear regression model, it should be noted that the moving horizon estimator (and, by analogy, the approximate nonlinear filters) can be put in the standard nonlinear regression form:

$$\begin{bmatrix} -X_{k-m+1|k-m} \\ y_{k-m+1} \\ 0 \\ 0 \\ y_{k-m+2} \\ \vdots \\ 0 \\ 0 \\ y_k \end{bmatrix} = \begin{bmatrix} -X_{k-m+1} \\ g(x_{k-m+1}, p_{k-m+1}) \\ F(x_{k-m+1}, p_{k-m+1}, u_{k-m+1}) - x_{k-m+2} \\ p_{k-m+1} - p_{k-m+2} \\ g(x_{k-m+2}, p_{k-m+2}) \\ \vdots \\ F(x_{k-1}, p_{k-1}, u_{k-1}) - x_k \\ p_{k-1} - p_k \\ g(x_k, p_k) \end{bmatrix} + \begin{bmatrix} X^e_{k-m+1} \\ v_{k-m+1} \\ w^x_{k-m+1} \\ w^p_{k-m+1} \\ v_{k-m+2} \\ \vdots \\ w^x_{k-1} \\ w^p_{k-1} \\ v_k \end{bmatrix}. \tag{48}$$

Although $F(\cdot)$ is an integral equation, recall that these problems are typically solved by approximating the ODEs by algebraic equations.

We have said that the EKF may produce biased estimates. Based on simulation results, Wishner et al. (1969) concluded that a second-order filter is less biased than the EKF or iterated linear-filter smoother (called single-stage iteration filter by them). To better understand the bias caused by nonlinearities, the reader is referred to Box (1971) who used a second-order Taylor series expansion to evaluate the bias in the weighted least-squares parameter estimates. (We note the similarity between his bias-correction term (Eq. 2.21 of Box) and the product of the gain and bias-correction term $(K_{TS}(t_i)\hat{b}_m(t_i^-))$ for the modified truncated second-order filter given in Eq. 12.44 of Maybeck (1982).)

Bates and Watts (1980) provide a great deal of insight into the adequacy of a first-order approximation of a nonlinear model. Their conclusion was that parameter transformations can reduce the effect of nonlinearities, resulting in a model that behaves more like a linear model. Ross (1990) lists several advantages of parameter transformations under the headings of "Numerical" (e.g., reduce the dependence on good initial guesses, fewer iterations until convergence), "Statistical" (e.g., reliability of estimates of confidence regions), and "Interpretational." Since all of the filtering methods discussed in this article are based on approximation of the conditional density, reducing the effect of nonlinearities can result in better performance (in the sense that the mean and variance estimates will tend to be a better approximation of the conditional density). Espie and Macchietto (1988) examined the numerical advantages of parameter transformations for several different nonlinear models. They found that appropriate transformations could significantly reduce the number of iterations required to solve a least-squares problem and the dependence on good initial guesses.

Optimization-based estimators

As stated in the Introduction, a number of optimization-based estimators have been proposed in recent years. Although the EKF is often used as a benchmark to compare their performance, the statistical equivalence between the two is usually ignored. Many published methods for parameter estimation are based on the following least-squares objective function (using our notation and a moving horizon formulation):

$$\min_{x_{k-m+1},\, p_{k-m+1}} \sum_{\ell=k-m+1}^{k} \nu_\ell^T R^{-1} \nu_\ell \tag{49}$$

s.t.

$$\frac{dx_t}{dt} = f(x_t, p, u_t)$$

$$y_\ell = g(x_\ell, p_\ell) + \nu_\ell .$$

This model is a special case of the dynamic model that we consider where all parameters are constant ($w_k^p = 0$, $\forall k$), the state equations are completely deterministic ($w_k^x = 0$, $\forall k$, which implies that initial state and parameters completely determine the values of all future states), and the prior distribution of the initial states and parameters is uniform ($P_{k-m+1|k-m}^{-1} = 0$). It will often be the case that time-varying parameters, unmeasured disturbances, or model errors in the state equations prohibit the use of such a model (Kozub and MacGregor, 1992).

When the measurement sequence $\{y_{k-m+1}, \dots , y_k\}$ contains enough information to obtain suitable estimates of the states and parameters, the assumption of uniformly distributed states and parameters ($P_{k-m+1|k-m}^{-1} = 0$) is a reasonable choice. However, for on-line applications, low signal-to-noise ratios, or nonlinear systems, there is motivation to reduce the computational requirements (i.e., horizon size) by weighting the initial estimates.

To track time-varying parameters this method can be implemented with a moving measurement horizon (with the parameters assumed to be constant over the horizon). However, it is necessary to find a compromise between choosing a small enough horizon size so that *all* parameters can be assumed constant and errors in the state equations are not significant and a large enough horizon size so that the measurements contain enough information to identify the parameters and states. Often, such a compromise does not exist. Furthermore, there is no way to model parameters that change at different time scales.

Ramamurthi et al. (1993) proposed using linearized state equations based on a predetermined reference trajectory for the objective function of Eq. 49. Their objective is a special case of the linear least squares formulation in this article. Error-in-variables methods (see, for example, Reilly and Patino-Leal, 1981; Ricker, 1984) assume measurement error in the independent process variables. In the preceding optimization problem, this would require estimation of the "true" process inputs. In this case, an additional error term corresponding to the measurement error in the input u (analogous to the measurement noise ν for the output y) should be added to the objective function. Kim et al. (1991) and Liebman et al. (1992) proposed an objective function similar to Eq. 49 for the error-in-variables case.

Example 1: *Robustness to Outliers.* In practice, it is very difficult to determine the precise values of the covariances for external noises. In fact, external signals are often nonstationary and their covariance changes in an unpredictable manner. When the Kalman filter is designed based on constant covariances, sudden changes in the covariance can lead to unrealistic state estimates. The ability of the least-squares estimator to use *a priori* known limits on the unmeasured disturbances, measurement noise, and states can potentially improve the reliability of the estimates in this instance. We demonstrate this fact with a simple example: a measurement outlier in a single input, single output (SISO) system.

The following SISO model was used for simulation:

$$x(s) = \frac{3.8}{15s+1} w(s)$$

$$y(s) = x(s) + \nu(s), \tag{50}$$

where x is the state to be estimated, y is a measurement of x with unit sampling time and corrupted with noise ν, and w is an unmeasured disturbance that is truncated normal on the interval $(-0.2, 0.2)$ with mean zero and covariance of 0.0064. The measurement noise ν has covariance 0.01. The initial state is zero and the initial estimate was generated from a normal distribution with covariance 10^{-6}. At time 10 the measured output y is 2. This represents an outlier in the measurement sequence. When the known bounds on the disturbance w are used as constraints for the least-squares estimation, the effect of the outlier is reduced. This is due to the fact that even the largest possible disturbance ($w = 0.2$) could not have caused such a rapid change in the output, so the measurement must be corrupted by noise. Results are given in Figure 1. Admittedly, measurement outlier detection by itself may not justify the use of the on-line optimization method. We have included this example only to show that various practical issues in state estimation can be treated within the proposed methodology.

Example 2: *Nonlinear Parameter Estimation for a Batch Fermentation Reactor.* This example is based on a batch fermentation reactor model presented in Ghose and Ghosh (1976). The reaction system is the conversion of glucose to gluconic acid. The state equations are

$$\frac{dC_C}{dt} = \mu_m \frac{C_C C_G C_O}{K_s C_O + K_0 C_G + C_G C_O} \tag{51}$$

$$\frac{dC_L}{dt} = \upsilon_L \frac{C_C C_G}{K_L + C_G} - 0.9082 K_p C_L \tag{52}$$

$$\frac{dC_A}{dt} = K_p C_L \tag{53}$$

$$\frac{dC_G}{dt} = -\frac{1}{Y_s} \mu_m \frac{C_C C_G C_o}{K_s C_O + K_0 C_G + C_G C_O} - 1.011 \upsilon_L \frac{C_C C_G}{K_L + C_G} \tag{54}$$

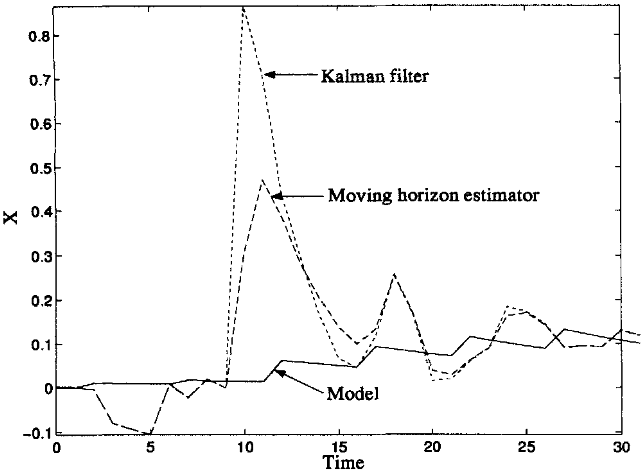

Figure 1. State estimation for Kalman filter and moving horizon estimator.

$$\frac{dC_O}{dt} = K_La(C_O^* - C_O) - \frac{1}{Y_0}\mu_m \frac{C_CC_GC_O}{K_sC_O + K_0C_G + C_GC_O} - 0.09v_L\frac{C_CC_G}{K_L + C_G}, \qquad (55)$$

where C_C is the cell concentration, C_L is gluconolactone concentration, C_A is gluconic acid concentration, C_G is glucose concentration, C_O is dissolved oxygen concentration, and C_O^* is the equilibrium concentration of dissolved oxygen. For this example it is assumed that the maximum specific growth rate, μ_m, and the velocity constant for lactone production, v_L, are unknown constants. The remaining model parameters are assumed known, and the interested reader can find their physical interpretations in Ghose and Ghosh (1976). Each batch run was simulated for 12 hours with measurements of C_C, C_A, and C_O, made at 10-minute intervals.

The initial conditions for the states and parameters are

$$\begin{bmatrix} C_C \\ C_L \\ C_A \\ C_G \\ C_O \end{bmatrix} = \begin{bmatrix} 0.1\ \mathrm{UOD/mL} \\ 0\ \mathrm{g/L} \\ 0\ \mathrm{g/L} \\ 50\ \mathrm{g/L} \\ 6.31\times10^{-3}\ g/L \end{bmatrix} \begin{bmatrix} \mu_m \\ v_L \end{bmatrix} = \begin{bmatrix} 0.39\ \mathrm{h}^{-1} \\ 8.30\ \mathrm{mg/UOD/h} \end{bmatrix}.$$

In order to make the estimation problem better conditioned, the states and parameters were normalized so that they remained (as much as possible) within an order of magnitude throughout the run. Also, the estimates for the states and parameters were required to be nonnegative.

The state vector is augmented by the unknown parameters, resulting in seven states to be estimated. The covariance matrices used in the simulation are

$$P_{1|0} = \mathrm{diag}\,[9\times10^{-4}\ 10^{-8}\ 10^{-8}\ 225\ 3.6\times10^{-6}\ 0.014\ 6.2] \qquad (56)$$

$$Q = \mathrm{diag}\,[10^{-6}\ 10^{-6}\ 10^{-6}\ 10^{-6}\ 10^{-6}\ 10^{-6} 10^{-6}] \qquad (57)$$

$$R = \mathrm{diag}\,[10^{-6}\ 10^{-4}\ 10^{-12}]. \qquad (58)$$

The elements of P were chosen so the initial estimates would have a standard deviation of 30% of the initial conditions. The elements of Q reflect the fact that there is little model error and the parameters are constant. The elements of R were chosen so that the respective normalized states would be measured with variance of 10^{-6}.

Twenty-five Monte Carlo simulations were made for the EKF and EKF-based moving horizon estimator with horizons 2, 3 and 10. For each run, the true values of the initial states and parameters remained constant, while the initial estimates and measurement noise were generated randomly from a ñormal distribution with the covariance matrices given earlier. The average sum of the squared errors (SSE) is given in Table 1 for the normalized state and parameter estimates. Of the 25 runs, the EKF failed to converge to the true state/parameter estimates on three runs. The moving horizon estimator converged on all runs for all horizon sizes. A typical run is presented in Figures 2 and 3, and a run where the EKF did not converge is presented in Figures 4 and 5.

For all runs the EKF and the moving horizon estimator track the measured states well. The difficulty in this example is the estimation of the unmeasured glucose state and the unknown model parameters. The glucose state appears nonlinearly in the kinetic equations and also multiplies the unknown parameters. Hence, the use of a linearized model to compute the estimates of these variables can lead to significant errors. The advantages of the moving horizon estimator compared to the EKF discussed before enable it to track these unmeasured parameters more reliably.

Conclusions

In this article, we presented a least squares formulation of the general state estimation problem that offers the advantages of being able to incorporate nonlinear models and inequality constraints on the estimated variables. The probabilistic interpretation provided a theoretical justification of the technique and insights into the choice of weighting matrices. In order to prevent the computational requirements from becoming unwieldy, a moving estimation horizon similar to the moving control horizon used in MPC was proposed. This provided the user with a convenient method to trade off the

Table 1. Sum of the Squared Errors for the State and Parameter Estimates

Method	C_C ($\times10^{-4}$)	C_L ($\times10^{-1}$)	C_A ($\times10^{-5}$)	C_G ($\times10$)	C_O ($\times10^{-4}$)	μ_m	v_L ($\times10$)
EKF	158.0	390.8	168.0	266.3	518.8	16.7	6.9
Moving horizon estimator							
$m = 2$	4.4	2.0	4.4	3.3	1.8	2.8	1.9
$m = 3$	3.2	1.4	4.5	4.1	2.1	2.6	1.8
$m = 10$	1.1	0.16	2.9	4.5	0.87	2.4	2.0

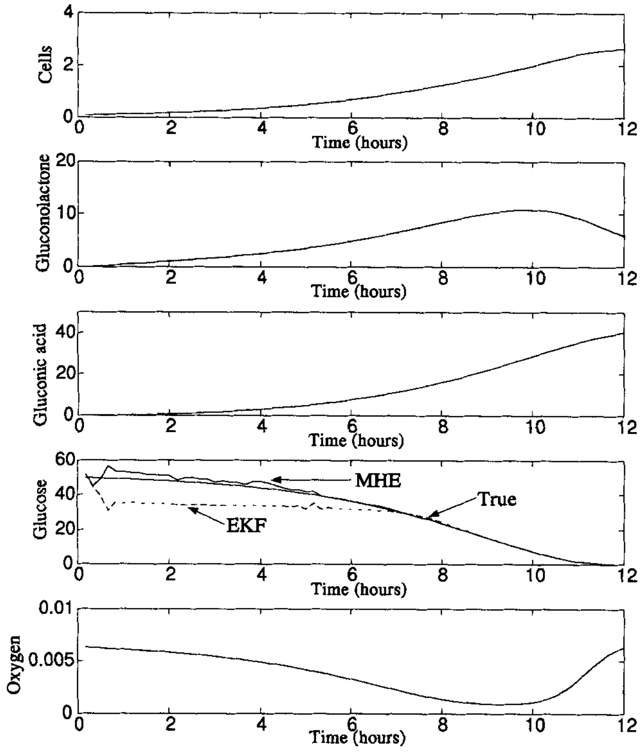

Figure 2. State estimates for the EKF and moving horizon estimator ($m = 2$) for the batch fermentation example.

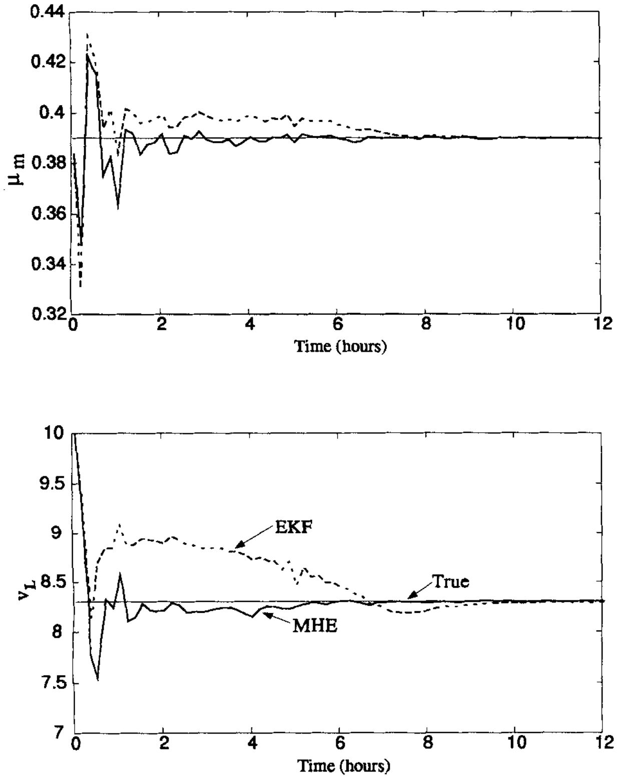

Figure 3. Parameter estimates for the EKF and moving horizon estimator ($m = 2$) for the batch fermentation example.

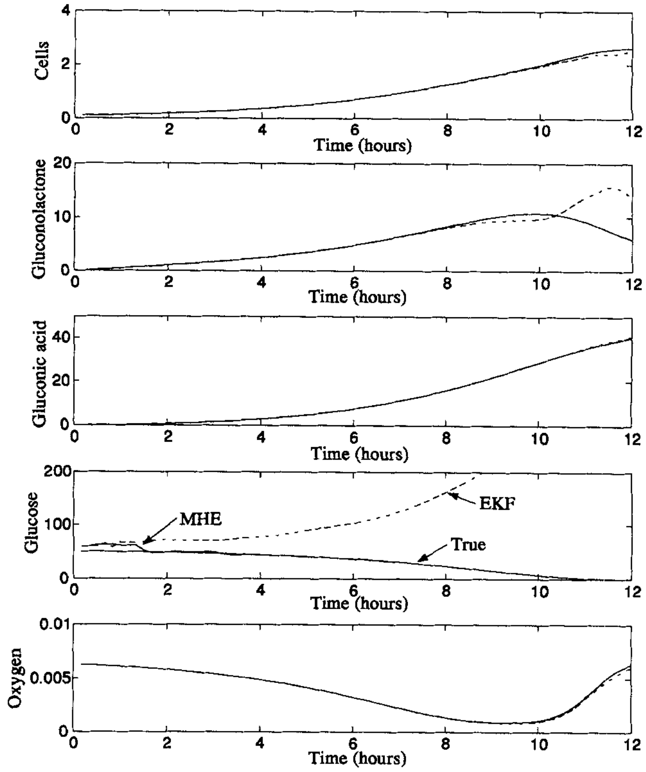

Figure 4. State estimates for the EKF and moving horizon estimator ($m = 2$) for the batch fermentation example.

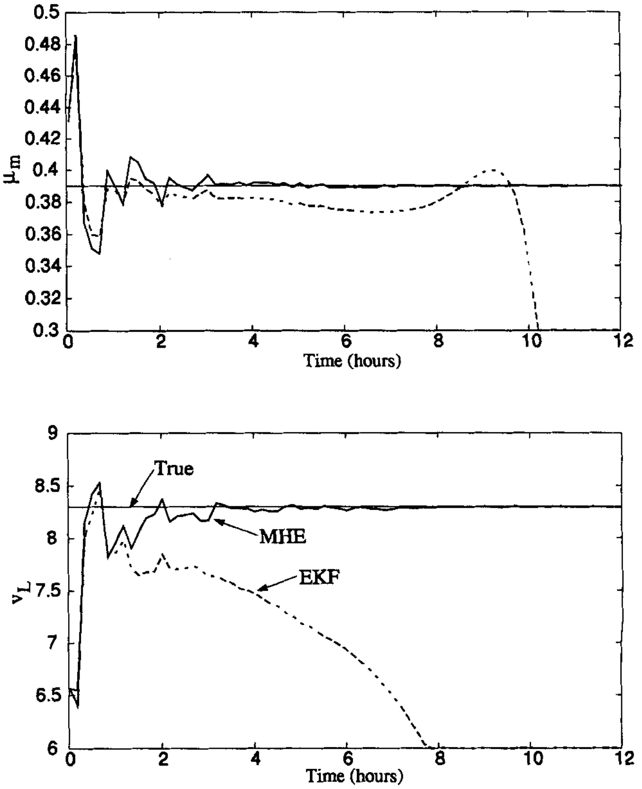

Figure 5. Parameter estimates for the EKF and moving horizon estimator ($m = 2$) for the batch fermentation example.

estimation accuracy against computational efficiency. The equivalence between the moving horizon estimator and well-known stochastic estimation techniques, like the Kalman filter, extended Kalman filter, and other approximate nonlinear filters, was discussed.

Based on this discussion, the following approach to nonlinear estimation is recommended. Whenever possible, make use of parameter transformations to reduce the effects of nonlinearities. Use as large a horizon size as computationally feasible. As the horizon size increases, the influence of the initial state and weighting matrix is reduced. When the measurement sequence $\{y_{k-m+1}, \ldots, y_k\}$ is not expected to contain enough information to adequately identify the states and parameters, the initial state and weighting matrix will have an important effect on the estimates, and extra effort should be made to ensure that $X_{k-m+1|k-m}$ and $P_{k-m+1|k-m}$ provide a reasonable description of past information.

Acknowledgment

One of the authors (J.H.L.) gratefully acknowledges financial support from the National Science Foundation (under contracts NSF-CTS-9209808 and NSF-CTS-9357827).

Notation

A = state transition matrix for linearized model
B = input matrix for linearized model
P = covariance of augmented state/parameter estimates
Q = covariance of random noise in the augmented state/parameter equations
R = covariance of random noise in the measurement equations; ideal gas constant
X^e = error in the initial state estimate
$f(\cdot)$ = state equations
$g(\cdot)$ = measurement equations
Γ = input matrix for linear model
Φ = augmented state/parameter transition matrix for linear and linearized model
Ξ = measurement matrix for linear and augmented linearized model

Literature Cited

Albert, A., and R. W. Sittler, "A Method for Computing Least Squares Estimators that Keep Up with the Data," *SIAM J. Contr.*, **3**, 384 (1966).

Albuquerque, J. A., and L. T. Biegler, "Decomposition Algorithms for On-Line Estimation with Nonlinear Models,"*Comp. Chem. Eng.* (1994).

Bates, D. M., and D. G. Watts, "Relative Curvature Measures of Nonlinearity," *J. R. Stat. Soc. Ser. B*, **42**, 1 (1980).

Box, M. J., "Bias in Nonlinear Estimation," *J. R. Stat. Soc. Ser. B*, **33**, 171 (1971).

Cox, H., "On the Estimation of State Variables and Parameters for Noisy Dynamic Systems," *IEEE Trans. Automat. Contr.*, **AC-9**, 5 (1964).

Cuthrell, J. E., and L. T. Biegler, "On the Optimization of Differential–Algebraic Process Systems," *AIChE J.*, **33**, 1257 (1987).

Espie, D. M., and S. Macchietto, "Nonlinear Transformations for Parameter Estimation," *Ind. Eng. Chem. Res.*, **27**, 2175 (1988).

Friedland, B., and I. Bernstein, "Estimation of the State of a Nonlinear Process in the Presence of Nongaussian Noise and Disturbances," *J. Franklin Inst.*, **281**, 455 (1966).

Gelb, A., ed., *Applied Optimal Estimation*, M.I.T. Press, Cambridge, MA (1974).

Ghose, T. K., and P. Ghosh, "Kinetic Analysis of Gluconic Acid Production by *Pseudomonas Ovalis*," *J. Appl. Chem. Biotechnol.*, **26**, 768 (1976).

Gill, P. E., W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, London (1981).

Ho, Y. C., and R. C. K. Lee, "A Bayesian Approach to Problems in Stochastic Estimation and Control," *IEEE Trans. Automat. Contr.*, **AC-9**, 333 (1964).

Hsia, T. C., *System Identification: Least Squares Methods*, Lexington Books, Lexington, MA (1977).

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).

Kalman, R. E., "A New Approach to Linear Filtering and Prediction Problems," *Trans. ASME, Ser. D: J. Basic Eng.*, **82**, 35 (1960).

Kalman, R. E., and R. S. Bucy, "New Results in Linear Filtering and Prediction Theory," *Trans. ASME, Ser. D: J. Basic Eng.*, **83**, 95 (1961).

Kim, I.-W., M. J. Liebman, and T. F. Edgar, "A Sequential Error-in-Variables Method for Nonlinear Dynamic Systems," *Comp. Chem. Eng.*, **15**, 663 (1991).

Kopp, R. E., and R. J. Oxford, "Linear Regression Applied to System Identification and Adaptive Control Systems," *AIAA J.*, **1**, 2300 (1963).

Kozub, D. J., and J. F. MacGregor, "State Estimation for Semi-Batch Polymerization Reactors," *Chem. Eng. Sci.*, **47**, 1047 (1992).

Kushner, H. J., "On the Differential Equations Satisfied by Conditional Probability Densities of Markov Processes," *SIAM J. Contr.*, **2**, 106 (1964).

Liebman, M. J., T. F. Edgar, and L. S. Lasdon, "Efficient Data Reconciliation and Estimation for Dynamic Processes Using Nonlinear Programming Techniques," *Comp. Chem. Eng.*, **16**, 963 (1992).

Ljung, L., "Asymptotic Behavior of the Extended Kalman Filter as a Parameter Estimator for Linear Systems," *IEEE Trans. Automat. Contr.*, **AC-24**, 36 (1979).

Ljung, L., "Recursive Identification Methods for Off-Line Identification Problems," *IFAC Symp. Ident. System Param. Est.*, Washington, D.C., p. 555 (1982).

Maybeck, P. S., *Stochastic Models, Estimation, and Control*, Vol. 2, Academic Press, New York (1982).

Muske, K. R., and J. B. Rawlings, "Model Predictive Control with Linear Models," *AIChE J.*, **39**, 262 (1993a).

Muske, K. R., J. B. Rawlings, and J. H. Lee, "Receding Horizon Recursive State Estimation," *Proc. 1993 American Control Conf.*, p. 900 (1993).

Rawlings, J. B., E. S. Meadows, and K. R. Muske, "Nonlinear Model Predictive Control: A Tutorial and Survey," *Proc. 1994 ADCHEM Conf.*, Kyoto, Japan (1994).

Reilly, P. M., and H. Patino-Leal, "A Bayesian Study of the Error-in-Variables Model," *Technometrics*, **23**, 221 (1981).

Ricker, N. L., "Comparison of Methods for Nonlinear Parameter Estimation," *Ind. Eng. Chem. Process Des. Dev.*, **23**, 283 (1984).

Ross, G. J. S., *Nonlinear Estimation*, Springer-Verlag, New York (1990).

Villadsen, J., and M. L. Michelsen, *Solution of Differential Equation Models by Polynomial Approximation*, Prentice Hall, Englewood Cliffs, NJ (1978).

Wishner, R. P., J. A. Tabaczynski, and M. Athans, "A Comparison of Three Non-Linear Filters," *Automatica*, **5**, 487 (1969).

Appendix

The proof assumes that the reader is familiar with the corresponding proof for the unconstrained case (e.g., see Jazwinski, pp. 153–154, 1970). To demonstrate the concepts while avoiding cumbersome notation, we assume that the system can be described by the following linear model:

$$x_{k+1} = Ax_k + w_k \tag{A1}$$

$$y_k = x_k + v_k, \tag{A2}$$

where $v_k \sim N(0, R)$ and w_k are mutually uncorrelated. The extension to the nonlinear model considered in this article is straightforward (the proof in Jazwinski (1970) for the unconstrained case is in terms of a nonlinear model).

Using Bayes' rule, we can write the conditional density of the states as

$$p(x_1, \dots, x_n \mid y_1, \dots, y_n) = \frac{p(y_1, \dots, y_n \mid x_1, \dots, x_n) p(x_1, \dots, x_n)}{p(y_1, \dots, y_n)}. \quad \text{(A3)}$$

Constraints on the random noise terms

To simplify the proof, (1) assume that only the $\{w_k\}$ are constrained (the extension to constraints on x_1 or $\{v_k\}$ is trivial), and (2) assume that the constraints are time invariant (again, extension to time-varying constraints is trivial):

$$w_{\min} \le w_k \le w_{\max} \qquad 1 \le k \le n-1. \quad \text{(A4)}$$

Define an indicator function for the set $\mathcal{A}$ as

$$I_{\mathcal{A}}(s) = \begin{cases} 1 & s \in \mathcal{A} \\ 0 & s \notin \mathcal{A} \end{cases}. \quad \text{(A5)}$$

Based on the constraints of Eq. A4, we can model w_k as a truncated normal random variable with density function:

$$p(w_k) = c \exp\left(-\frac{1}{2} w_k^T Q^{-1} w_k \right) \cdot I_{[w_{\min}, w_{\max}]}(w_k). \quad \text{(A6)}$$

Note that the density is zero whenever w_k is outside the interval $[w_{\min}, w_{\max}]$. The normalizing constant c ensures that $p(w_k)$ is a proper density (i.e., its integral is one):

$$c = \frac{1}{\int_{w_{\min}}^{w_{\max}} \exp\left(-\frac{1}{2} W^T Q^{-1} W \right) dW}, \quad \text{(A7)}$$

where W is a variable of integration and does not correspond to the actual value of w_k in Eq. A6. Note that W is a vector, and the integration is over a region whose dimension is equal to that of w_k.

Constraints on w_k alter the prior distribution of the state trajectory $p(x_1, \dots, x_n)$ but do not affect the likelihood function $p(y_1, \dots, y_n \mid x_1, \dots, x_n)$; therefore, to prove the MAP interpretation, we must show that the exponential term of the prior density corresponds to that of the unconstrained case and that the normalizing constant does not depend on the estimates. Since $\{x_k\}$ is a Markov sequence

$$\begin{aligned} p(x_1, \dots, x_n) &= p(x_1) \prod_{k=2}^{n} p(x_k | x_{k-1}) \\ &= p(x_1) \prod_{k=2}^{n} p_w(x_k - Ax_{k-1}) \\ &= p(x_1) \prod_{k=2}^{n} c \exp\left[-\frac{1}{2}(x_k - Ax_{k-1})^T Q^{-1}(x_k - Ax_{k-1}) \right] \\ &\qquad \cdot I_{[w_{\min}, w_{\max}]}(x_k - Ax_{k-1}) \\ &= p(x_1) c^{n-1} \exp\left[-\frac{1}{2} \sum_{k=2}^{n} (x_k - Ax_{k-1})^T Q^{-1}(x_k - Ax_{k-1}) \right] \\ &\qquad \cdot I_{\mathcal{B}}(x_2 - Ax_1, \dots, x_n - Ax_{n-1}) \\ &= p(x_1) c^{n-1} \exp\left(-\frac{1}{2} \sum_{k=1}^{n-1} w_k^T Q^{-1} w_k \right) \cdot I_{\mathcal{B}}(w_1, \dots, w_{n-1}), \end{aligned} \quad \text{(A8)}$$

where $\mathcal{B}$ is the cuboid $[w_{\min}, w_{\max}] \times \dots \times [w_{\min}, w_{\max}]$ in the parameter space of $\{w_k\}$. Note that $p(x_1, \dots, x_n) = 0$ whenever $\{w_1, \dots, w_k\} \in \mathcal{B}$ and $p(x_1, \dots, x_n) > 0$ whenever $\{w_1, \dots, w_{\mathrm{k}}\} \in \mathcal{B}$. This implies that the values of $\{w_k\}$ that maximize the posterior density, Eq. A3, must satisfy the constraints, Eq. A4. The additional fact that c does not depend on the $\{w_k\}$ (only on $w_{\min}$ and $w_{\max}$, which are chosen in advance) and $I_{\mathcal{B}}(w_1, \dots, w_{n-1})$ is either zero (if a constraint is violated) or one, implies that minimizing the least-squares objective of Eq. 10 subject to Eq. A4 is equivalent to maximizing the conditional density of the states.

Constraints on the states

To simplify the presentation, assume that a time-invariant upper bound is known for the states:

$$x_k \le x_{\max} \qquad 2 \le k \le n. \quad \text{(A9)}$$

The prior distribution of the state trajectory is determined by the distribution of $\{w_k\}$. To understand how state constraints affect the distribution of $\{w_k\}$, Eq. A9 can be rewritten as

$$A^k x_1 + \sum_{i=1}^{k} A^{k-i} w_i \le x_{\max} \qquad 1 \le k \le n-1. \quad \text{(A10)}$$

As before, only the prior distribution of the states is altered. However, in this case, there are several possible models for $\{w_k\}$ that satisfy the preceding constraints. We discuss two alternatives. The first is chosen so that the least-squares estimates maximize the conditional joint density; however, the underlying model is no longer Markov. The second results in a conditional density that cannot be maximized by the solution of a least-squares problem, but is of interest since it may be more appropriate for modeling physical processes. It is also interesting to understand when the solution of the least-squares objective is good approximation to the MAP estimate for the second model.

Model 1. The constraints of Eq. A10 construct a feasible region in the parameter space of $\{w_k\}$ (for nonlinear models this region is somewhat abstract, but does not invalidate the proof). If we assume that $\{w_k\}$ is jointly truncated normal over this region, the density can be written as

$$p(w_1, \dots, w_{n-1} \mid x_1) = c \prod_{k=1}^{n-1} \exp\left(-\frac{1}{2} w_k^T Q^{-1} w_k \right) \cdot I_{\mathcal{B}_k}(w_k), \quad \text{(A11)}$$

where $\mathcal{B}_k = [-\infty, x_{\max} - A^k x_1 - \Sigma_{i=1}^{k-1} \mathrm{A}^{\mathrm{k}-\mathrm{i}} w_i]$. The normalizing constant for this density is

$$c^{-1} = \int_{-\infty}^{x_{\max} - Ax_1} \int_{-\infty}^{x_{\max} - A^2 x_1 - AW_1} \cdots \int_{-\infty}^{x_{\max} - A^{n-1} x_1 - \Sigma_{i=1}^{n-2} A^{n-i-1} W_i} \exp\left(-\frac{1}{2} \sum_{i=1}^{n-1} W_i^T Q^{-1} W_i \right) dW_{n-1} \cdots dW_1, \quad \text{(A12)}$$

where the $\{W_k\}$ are variables of integration and do not correspond to the values of $\{w_k\}$ in the joint density (A11).

The derivation of the joint density of the states is different than that of (Eq. A8), since $\{w_k\}$ is no longer an independent sequence:

$$\begin{aligned}
p(x_1, \ldots, x_n) &= p(x_1) \prod_{k=2}^{n} p(x_k \mid x_1, \ldots, x_{k-1}) \\
&= p(x_1) \prod_{k=2}^{n} p_{w_{k-1} \mid x_1, w_1, \ldots, w_{k-2}}(x_k - Ax_{k-1} \mid x_1, x_2 \\
&\qquad - Ax_1, \ldots, x_{k-1} - Ax_{k-2}) \\
&= p(x_1) \prod_{k=2}^{n} p_{w_{k-1} \mid x_1, w_1, \ldots, w_{k-2}}(w_{k-1} \mid x_1, w_1, \ldots, w_{k-2}) \\
&= p(x_1) p_{w_1, \ldots, w_{n-1} \mid x_1}(w_1, \ldots, w_{n-1} \mid x_1) \\
&= p(x_1) c \prod_{k=1}^{n-1} \exp\left(-\frac{1}{2} w_k^T Q^{-1} w_k\right) \cdot I_{\mathcal{B}_k}(w_k) \\
&= p(x_1) c \exp\left(-\frac{1}{2} \sum_{k=1}^{n-1} w_k^T Q^{-1} w_k\right) \cdot I_{\mathcal{B}}(w_1, \ldots, w_{n-1}),
\end{aligned} \tag{A13}$$

where $\mathcal{B} = \mathcal{B}_1 \times \ldots \times \mathcal{B}_{n-1}$, with $\mathcal{B}_k$ defined as before. The expression, Eq. A13, is in the same form as Eq. A8. Since the region $\mathcal{B}$ encloses that parameter space of $\{w_k\}$ that satisfies the state constraints and c does not depend on the particular realization of $\{w_k\}$, a similar argument shows that minimizing the least-squares objective, Eq. 10 subject to Eq. A9 with the joint density of $\{w_k\}$ modeled as in Eq. A11 is equivalent to maximizing the conditional density of the states. Note that the joint distribution, Eq. A11, assumed for the $\{w_k\}$ resulted in a state sequence that is no longer Markov (compare the first equality in Eq. A8 with the first equality in Eq. A13).

Model 2. A second possible model is to assume that, given x_k, w_k is a truncated normal random variable over the range of values that do not violate the state constraints on x_{k+1}. In this case, define a truncated random variable w_k whose density satisfies the preceding constraints:

$$p(w_k \mid x_k) = c_k \exp\left(-\frac{1}{2} w_k^T Q^{-1} w_k\right) \cdot I_{\mathcal{B}_k}(w_k), \tag{A14}$$

where $\mathcal{B}_k = [-\infty, x_{\max} - Ax_k] = [-\infty, x_{\max} - A^k x_1 - \Sigma_{i=1}^{k-1} A^{k-i} w_i]$. The normalizing constant for this density is

$$c_k^{-1} = \int_{-\infty}^{x_{\max} - A^k x_1 - \Sigma_{i=1}^{k-1} A^{k-i} w_i} \exp\left(-\frac{1}{2} W^T Q^{-1} W\right) dW. \tag{A15}$$

Notice that the area of integration for c_k depends on $\{w_1, \ldots, w_{k-1}\}$. The joint density of the state becomes

$$\begin{aligned}
p(x_1, \ldots, x_n) &= p(x_1) \prod_{k=2}^{n} p(x_k \mid x_{k-1}) \\
&= p(x_1) \prod_{k=2}^{n} p_{w_{k-1} \mid x_{k-1}}(x_k - Ax_{k-1} \mid x_{k-1}) \\
&= p(x_1) \prod_{k=2}^{n} c_{k-1} \exp\left[-\frac{1}{2}(x_k - Ax_{k-1})^T Q^{-1}(x_k - Ax_{k-1})\right] \cdot I_{\mathcal{B}_{k-1}}(x_k - Ax_{k-1}) \\
&= p(x_1) c_1 c_2 \cdots c_{n-1} \exp\left[-\frac{1}{2} \sum_{k=2}^{n} (x_k - Ax_{k-1})^T Q^{-1}(x_k - Ax_{k-1})\right] \cdot I_{\mathcal{B}}(x_2 - Ax_1, \ldots, x_n - Ax_{n-1}) \\
&= p(x_1) c_1 c_2 \ldots c_{n-1} \exp\left(-\frac{1}{2} \sum_{k=1}^{n-1} w_k^T Q^{-1} w_k\right) \cdot I_{\mathcal{B}}(w_1, \ldots, w_{n-1}),
\end{aligned} \tag{A16}$$

where $\mathcal{B} = \mathcal{B}_1 \times \ldots \times \mathcal{B}_{n-1}$, with $\mathcal{B}_k$ defined as before. Since the normalizing constants in Eq. A16 are functions of the $\{w_k\}$, maximizing the exponential term does not necessarily maximize the entire expression; therefore, the solution of the least-squares problem, Eq. 10, does not necessarily maximize the joint conditional density of the states.

Manuscript received Oct. 10, 1995, and revision received Dec. 11, 1995.