New Finite-Dimensional Filters for Parameter Estimation of Discrete-Time Linear Gaussian Models

Robert J. Elliott and Vikram Krishnamurthy, *Member, IEEE*

Abstract— **In this paper the authors derive a new class of finite-dimensional recursive filters for linear dynamical systems. The Kalman filter is a special case of their general filter. Apart from being of mathematical interest, these new finite-dimensional filters can be used with the expectation maximization (EM) algorithm to yield maximum likelihood estimates of the parameters of a linear dynamical system. Important advantages of their filter-based EM algorithm compared with the standard smoother-based EM algorithm include: 1) substantially reduced memory requirements and 2) ease of parallel implementation on a multiprocessor system. The algorithm has applications in multisensor signal enhancement of speech signals and also econometric modeling.**

Index Terms— **Expectation maximization algorithm, finite-dimensional filters, Kalman filter, maximum likelihood parameter estimation.**

I. Introduction

There are very few estimation problems for which finite-dimensional optimal filters exist, i.e., filters given in terms of finite-dimensional sufficient statistics. Indeed the only two cases that are widely used are the Kalman filter for linear Gaussian models and the Wonham filter (hidden Markov model filter) for finite state Markov chains in white noise.

In this paper we derive new finite-dimensional filters for linear Gaussian state-space models in discrete-time. The filters compute all the statistics required to obtain maximum likelihood estimates (MLE's) of the model parameters via the expectation maximization (EM) algorithm. The Kalman filter is a special case of these general filters.

MLE's of linear Gaussian models and other related time-series models using the EM algorithm were studied in the 1980's in [1] and [2] and more recently in the electrical engineering literature in [4] and [5]. The EM algorithm is a general iterative numerical algorithm for computing the MLE. Each iteration consists of two steps: the expectation (E-step) and the maximization (M-step). The E-step for linear Gaussian models involves computing the following two conditional expectations based on all the observations:

1) the sum over time of the state;
2) the sum over time of the state covariance.

In all the existing literature on parameter estimation of linear Gaussian models via the EM algorithm, the E-step is noncausal involving fixed-interval smoothing via a Kalman smoother (i.e., a forward pass and a backward pass).

In this paper we derive a *filter*-based EM algorithm for linear Gaussian models. That is, the E-step is implemented using filters (i.e., only a forward pass) rather than smoothers. The main contribution of this paper is to show that these filters are *finite-dimensional*. Few finite-dimensional filters are known, so the result is of interest.

It is important to note that the filter-based EM algorithm proposed here and the standard smoother-based EM algorithm in [1], [2], [4], and [5] are off-line iterative algorithms. They represent two different ways of computing the same conditional expectations and consequently yield the same result. However, the filter-based EM algorithm has the following advantages.

1) The memory costs are significantly reduced compared to the standard (smoother-based) EM algorithm.
2) The filters are decoupled and hence easy to implement in parallel on a multiprocessor system.
3) The filter-based EM algorithm is at least twice as fast as the standard smoother-based EM algorithm because no forward–backward scheme is required.

Filter-based EM algorithms have recently been proposed for hidden Markov models (HMM's) in [9]. These HMM filters are finite-dimensional because of the idempotent property of the state indicator function of a finite state Markov chain. In linear Gaussian models, unlike the HMM case, the state indicator vector is no longer idempotent. Instead, the filters derived in this paper are finite dimensional because of the following two algebraic properties that hold at each time instant.

1) The filtered density of the current time sum of the state is given by an affine function in x times the filtered state density. The filtered state density is a Gaussian in x with mean and variance given by the Kalman filter equations.
2) The filtered density of the current time sum of the state covariance is a quadratic in x times the filtered state density.

So the filtered density of the state sum is given in terms of four sufficient statistics, namely the two coefficients of the

Manuscript received August 8, 1997; revised June 9, 1998. Recommended by Associate Editor, T. Katayama. The work of R. Elliott was supported by NSERC under Grant A7964, the University of Melbourne, and the Co-operative Research Center for Sensor Signal and Information Processing in March 1995. V. Krishnamurthy was supported in part by an ARC grant, an ATERB grant, and the Co-operative Research Center for Sensor Signal and Information Processing.

R. J. Elliott is with the Department of Mathematical Sciences, University of Alberta, Edmonton, T6G 2G1 Canada.

V. Krishnamurthy is with the Department of Electrical and Electronic Engineering, University of Melbourne, Parkville, Victoria 3052 Australia (e-mail: vikram@ee.mu.oz.au.)

Publisher Item Identifier S 0018-9286(99)02823-8.

affine function in x and the Kalman mean and covariance. Similarly, the filtered density of the covariance sum is given by five sufficient statistics.

Actually this algebraic "closure" property holds for higher order statistics as well: We prove that the filtered density of the current time sum of the pth-order statistic of the state is a pth-order polynomial in x times the filtered state estimate. So finite-dimensional filters can be derived for the time sum of pth-order statistics of the state. Of course, for the filtered E-step we only use filters for the first and second order statistics. Also for $p = 0$, the filters reduce to the Kalman filter.

Applications: The filter-based EM algorithm proposed in this paper for linear Gaussian models can be applied to all the applications where the standard EM algorithm has been applied. In particular these include:

- multisensor signal enhancement algorithms for estimation of speech signals in room acoustic environments [4];
- high-resolution localization of narrowband sources using multiple sensors and direction of arrival estimation [7];
- linear predictive coding of speech (see [6, Ch. 6]);
- forecasting and prediction of the "shadow economy" in market cycles using linear errors-in-variables models [2].

In all these applications the advantages of the filter-based EM algorithm can be exploited.

This paper is organized as follows: In Section II we present the EM algorithm for maximum likelihood estimation of the parameters of a state-space linear Gaussian model to illustrate the use of the finite-dimensional filters derived in this paper. In Section III, a measure change is given which facilitates easy derivation of the filters. In Section IV, recursions are derived for the filtered densities of the variables of interest. In Section V, we derive the finite-dimensional filters. In Section VI a general finite-dimensional filter is proposed. Section VII re-expresses the filters to allow for singular state noise as long as a certain controllability condition is satisfied. In Section VIII an example of the filter-based EM algorithm for errors-in-variables time-series is given. In Section IX we evaluate the computational complexity of the filters and propose a parallel implementation. Finally conclusions are presented in Section X.

II. MLE of Gaussian State-Space Models

The aim of this section is to show how the finite-dimensional filters derived in this paper arise in computing the maximum likelihood parameter estimate of a linear Gaussian state-space model via the EM algorithm. We first briefly review the EM algorithm and describe the linear Gaussian state-space model. The use of the EM algorithm to compute maximum likelihood parameter estimates of the Gaussian state-space model is then illustrated. Finally, the use of the finite-dimensional filters to implement the filter-based EM algorithm is demonstrated. This motivates the finite-dimensional filters derived in the rest of the paper.

A. Review of the EM Algorithm

The EM algorithm is a widely used iterative numerical algorithm for computing maximum likelihood parameter estimates of partially observed models such as linear Gaussian state-space models, e.g., [2], [5] and HMM's [11]. For such models, direct computation of the MLE is difficult. The EM algorithm has the appealing property that successive iterations yield parameter estimates with nondecreasing values of the likelihood function.

Suppose we have observations $\{y_1, \cdots, y_T\}$ available, where T is a fixed positive integer. Let $\{P_\theta, \theta \in \Theta\}$ be a family of probability measures on $(\Omega, \mathcal{F})$ all absolutely continuous with respect to a fixed probability measure P_0. The likelihood function for computing an estimate of the parameter θ based on the information available in $\mathcal{Y}_T$ is

$$\mathcal{L}_T(\theta) = \mathbf{E}_0\left[\frac{dP_\theta}{dP_0}\,\middle|\,\mathcal{Y}_T\right]$$

and the MLE is defined by

$$\hat{\theta} \in \operatorname*{argmax}_{\theta\in\Theta} \mathcal{L}_T(\theta).$$

The EM algorithm is an iterative numerical method for computing the MLE. Let $\hat{\theta}_0$ be the initial parameter estimate. The EM algorithm generates a sequence of parameter estimates $\{\theta_j\}$, $j \in Z^+$, as follows.

Each iteration of the EM algorithm consists of two steps.

Step 1) (E-step) Set $\tilde{\theta} = \hat{\theta}_j$ and compute $Q(\cdot, \tilde{\theta})$, where

$$Q(\theta, \tilde{\theta}) = \mathbf{E}_{\tilde{\theta}}\left[\log \frac{dP_\theta}{dP_{\tilde{\theta}}}\,\middle|\,\mathcal{Y}_T\right].$$

Step 2) (M-step) Find $\hat{\theta}_{j+1} \in \operatorname{argmax}_{\theta\in\Theta} Q(\theta, \theta_j)$.

Using Jensen's inequality it can be shown (see [13, Th. 1]) that the sequence of model estimates $\{\hat{\theta}_j\}$, $j \in Z^+$ from the EM algorithm are such that the sequence of likelihoods $\{\mathcal{L}_T(\hat{\theta}_j)\}$, $j \in Z^+$ is monotonically increasing with equality if and only if $\hat{\theta}_{j+1} = \hat{\theta}_j$.

Sufficient conditions for convergence of the EM algorithm are given in [14]. We briefly summarize them and assume the following.

1) The parameter space Θ is a subset of some finite-dimensional Euclidean space R^r.
2) $\Omega_{\hat{\theta}_0} = \{\theta \in \Theta : \mathcal{L}_T(\theta) \geq \mathcal{L}_T(\hat{\theta}_0)\}$ is compact for any $\mathcal{L}_T(\hat{\theta}_0) > -\infty$.
3) $\mathcal{L}_T$ is continuous in Θ and differentiable in the interior of Θ. (As a consequence of 1)–3), clearly $\mathcal{L}_T(\hat{\theta}_j)$ is bounded from above).
4) The function $Q(\theta, \hat{\theta}_j)$ is continuous both in θ and $\hat{\theta}_j$.

Then by [14, Th. 2], the limit of the sequence of EM estimates $\{\hat{\theta}_j\}$ is a stationary point $\bar{\theta}$ of $\mathcal{L}_T$. Also $\{\mathcal{L}_T(\hat{\theta}_j)\}$ converges monotonically to $\bar{\mathcal{L}}_T = \mathcal{L}_T(\bar{\theta})$ for some stationary point $\bar{\theta}$. To make sure that $\bar{\mathcal{L}}_T$ is a maximum value of the likelihood, it is necessary to try different initial values $\hat{\theta}_0$.

B. Gaussian State-Space Model

To derive the filters with maximum generality, in this paper we consider a multi-input/multi-output linear Gaussian state-space model with time-varying parameters and noise variances as follows.

All processes are defined on the probability space $(\Omega, \mathcal{F}, P)$. We shall consider the classical linear-Gaussian model for the signal and observation processes. That is, for $k = 0, 1, \cdots$, assume that the state x_k is observed indirectly via the vector observations y_k, where

$$x_{k+1} = A_{k+1} x_k + B_{k+1} w_{k+1} \tag{1}$$

$$y_k = C_k x_k + D_k v_k. \tag{2}$$

Here x_k is a m-dimensional random vector. Also x_0 is a Gaussian random variable with zero mean and covariance matrix B_0^2 (of dimension $m \times m$).

At time $k = 1, 2, \cdots$, the noise in (1) is modeled by an independent Gaussian random variable with zero mean and covariance matrix B_k^2. It is known [8] that such a Gaussian random variable can be represented as $B_k w_k$ where w_k is an m-vector of independent $N(0,1)$ random variables.

In (2), for each k, $y_k \in R^d$ and v_k is a vector of independent $N(0,1)$ random variables. The process v_k is assumed to be independent of w_k. Assume that D_k is a nonsingular $d \times d$ matrix.

Finally, x_0 is assumed independent of the processes w_k and v_k.

Assumption 2.1: For the time being, assume that the $m \times m$ matrices B_k, $k = 0, 1, \cdots$, are nonsingular and symmetric. The case when B_k is singular is discussed in Section VII.

Remark: We assume B_k to be a covariance matrix and hence symmetric for notational convenience. The results in this paper also hold for *nonsymmetric* B_k, simply by replacing B_k^2 by $B_k B_k'$ and B_k^{-2} by $(B_k B_k')^{-1}$ below.

C. EM Algorithm for MLE of Gaussian State-Space Model

The aim of this subsection is to illustrate the use of the EM algorithm for computing the MLE of a Gaussian state-space model. Our main focus in this paper involves computation of the E-step. Hence, to keep the exposition simple, we omit issues of identifiability and consistency of the MLE. These are well known and can be found for example in [10, Ch. 7].

Consider the following time-invariant version of the linear Gaussian state-space model (1), (2)

$$x_{k+1} = A(\theta) x_k + B(\theta) w_{k+1} \tag{3}$$

$$y_k = C(\theta) x_k + D(\theta) v_k \tag{4}$$

where θ denotes the parameter vector belonging to some compact space Θ. Let $\theta^* \in \Theta$ denote the true model. We also assume other regularity conditions (see [10, Ch. 7]) on A, B, C, and D including identifiability of θ^* so that the MLE is strongly consistent, i.e., $\lim_{T \to \infty} \operatorname{argmax}_{\theta \in \Theta} \mathcal{L}_T(\theta)$ converges almost surely to the true model θ^*. For simplicity and in order to illustrate our main ideas, in this subsection we assume the parameterization $\theta = (A, B^2, C, DD')$. An errors-in-variables model with a different parameterization θ is given in Section VIII.

Suppose we wish to compute the MLE of the parameter $\theta = (A, B^2, C, DD')$ of (3), (4) given the observation sequence $y_0, \cdots, y_T$. We now illustrate the use of the EM algorithm outlined in Section II-A to compute the MLE of θ.

Step 1—E-Step: It is easily seen (see the Appendix or [4], [7], [2], or [1]) that for the model (3), (4)

$$\begin{aligned} Q(\theta, \hat{\theta}_j) &= -T \log|B| - (T+1)\log|D| \\ &\quad - \frac{1}{2} \mathbf{E}_{\hat{\theta}_j} \left\{ \sum_{l=1}^{T} (x_l - A x_{l-1})' B^{-2} (x_l - A x_{l-1}) \,\middle|\, \mathcal{Y}_T \right\} \\ &\quad - \frac{1}{2} \mathbf{E}_{\hat{\theta}_j} \left\{ \sum_{l=0}^{T} (y_l - C x_l)' (DD')^{-1} (y_l - C x_l) \,\middle|\, \mathcal{Y}_T \right\} \\ &\quad + \mathbf{E}_{\hat{\theta}_j} \{ R(\hat{\theta}_j) | \mathcal{Y}_T \} \end{aligned} \tag{5}$$

where $\hat{\theta}_j = (\hat{A}^{(j)}, \hat{B^2}^{(j)}, \hat{C}^{(j)}, \widehat{DD'}^{(j)})$ denotes the parameter estimate at the jth iteration and the term $R(\hat{\theta}_j)$ does not involve θ.

Step 2—M-Step: To implement the M-step, i.e., compute $\hat{\theta}_{j+1} = \operatorname{argmax}_{\theta \in \Theta} Q(\theta, \hat{\theta}_j)$, simply set the derivatives $\partial Q / \partial \theta = 0$. This yields (using the identity $\partial \log|M| / \partial M = M^{-1}$ for any nonsingular matrix M) the updated parameter estimate as $\hat{\theta}_{j+1} = (\hat{A}^{(j+1)}, \hat{B^2}^{(j+1)}, \hat{C}^{(j+1)}, \widehat{DD'}^{(j+1)})$ where

$$\begin{aligned} \hat{A}^{(j+1)} &= \mathbf{E}_{\hat{\theta}_j} \left\{ \sum_{l=1}^{T} x_l x_{l-1}' \,\middle|\, \mathcal{Y}_T \right\} \\ &\quad \times \left[\mathbf{E}_{\hat{\theta}_j} \left\{ \sum_{l=1}^{T} x_{l-1} x_{l-1}' \,\middle|\, \mathcal{Y}_T \right\} \right]^{-1} \\ \hat{B^2}^{(j+1)} &= \frac{1}{T} \mathbf{E}_{\hat{\theta}_j} \left\{ \sum_{l=1}^{T} \left(x_l - \hat{A}^{(j+1)} x_{l-1} \right) \right. \\ &\quad \left. \times \left(x_l - \hat{A}^{(j+1)} x_{l-1} \right)' \,\middle|\, \mathcal{Y}_T \right\} \\ \hat{C}^{(j+1)} &= \mathbf{E}_{\hat{\theta}_j} \left\{ \sum_{l=0}^{T} y_l x_l' \,\middle|\, \mathcal{Y}_T \right\} \left[\mathbf{E}_{\hat{\theta}_j} \left\{ \sum_{l=0}^{T} x_l x_l' \,\middle|\, \mathcal{Y}_T \right\} \right]^{-1} \\ \widehat{DD'}^{(j+1)} &= \frac{1}{T+1} \mathbf{E}_{\hat{\theta}_j} \left\{ \sum_{l=0}^{T} \left(y_l - \hat{C}^{(j+1)} x_l \right) \right. \\ &\quad \left. \times \left(y_l - \hat{C}^{(j+1)} x_l \right)' \,\middle|\, \mathcal{Y}_T \right\}. \end{aligned} \tag{6}$$

The above system (6) gives the EM parameter estimates at each iteration for the linear Gaussian model (3), (4). System (6) is well known. Indeed, versions of (6) with different parameterizations have appeared in several papers, e.g., [1], [2], [4], and [7]. Furthermore, since $Q(\theta, \theta_j)$ in (5) is continuous in θ and θ_j, as mentioned in Section II-A, the EM algorithm converges to a stationary point in the likelihood surface—see [2] for details.

Our main focus in this paper is the computation of the various conditional expectations in (5), and hence (6), which are required for implementing each iteration of the EM algorithm. It is well known that these conditional expectations can be computed via a Kalman smoother. Such an approach is termed the *smoother-based EM algorithm* and is described in

[1], [2], [4], and [7]. For example, consider the computation of $\mathbf{E}_{\hat{\theta}_j}\{\sum_{l=1}^{T} x_l x_l' \mid \mathcal{Y}_T\}$ in (5) and (6). Defining $\hat{x}_l = \mathbf{E}_{\hat{\theta}_j}\{x_l \mid \mathcal{Y}_T\}$, we have

$$\begin{aligned}\mathbf{E}_{\hat{\theta}_j}\left\{\sum_{l=1}^{T} x_l x_l' \,\middle|\, \mathcal{Y}_T\right\} &= \sum_{l=1}^{T} \mathbf{E}_{\hat{\theta}_j}\{x_l x_l' \mid \mathcal{Y}_T\} \\ &= \sum_{l=1}^{T} \mathbf{E}_{\hat{\theta}_j}\{(x_l - \hat{x}_l)(x_l - \hat{x}_l)' \mid \mathcal{Y}_T\} \\ &\quad + \hat{x}_l \hat{x}_l'.\end{aligned}$$

Now $\hat{x}_l$ and $\mathbf{E}_{\hat{\theta}_j}\{(x_l - \hat{x}_l)(x_l - \hat{x}_l)' \mid \mathcal{Y}_T\}$ are merely the smoothed state and covariance estimates computed via a fixed-interval Kalman smoother.

D. Filter-Based EM Algorithm

The main contribution of this paper is to show how the various conditional expectations in (5), and hence (6), can be computed using *causal filters* instead of smoothers. Thus we derive a *filter-based EM algorithm*. For example, we derive finite-dimensional filters for $H_k^0 \triangleq \sum_{l=1}^{k} x_l x_l'$, i.e., recursions for $\mathbf{E}_{\hat{\theta}_j}\{H_k^{(0)} \mid \mathcal{Y}_k\} = \mathbf{E}_{\hat{\theta}_j}\{\sum_{l=1}^{k} x_l x_l' \mid \mathcal{Y}_k\}$. Clearly at time $k = T$, the filtered estimate $\mathbf{E}_{\hat{\theta}_j}\{H_T^{(0)} \mid \mathcal{Y}_T\}$ is exactly what is required in the EM algorithm. Note that both the filter-based and smoother-based EM algorithms compute the same quantities and are off-line iterative algorithms. However, the filter-based EM algorithm has several advantages over the smoother-based EM algorithm as mentioned in Section I.

More specifically, defining the matrices

$$\begin{aligned} H_k^{(0)} &= \sum_{l=0}^{k} x_l x_l' \quad & H_k^{(1)} &= \sum_{l=1}^{k} x_l x_{l-1}' \\ H_k^{(2)} &= \sum_{l=1}^{k} x_{l-1} x_{l-1}' \quad & J_k &= \sum_{l=0}^{k} x_l y_l' \end{aligned} \tag{7}$$

the EM parameter estimates (6) at the $(j+1)$th iteration can be re-expressed as

$$\begin{aligned} \hat{A}^{(j+1)} &= \hat{H}_T^{(1)}\big(\hat{H}_T^{(2)}\big)^{-1}, \qquad \hat{C}^{(j+1)} = \hat{J}_T'\big(\hat{H}_T^{(0)}\big)^{-1} \\ \hat{B^2}^{(j+1)} &= \frac{1}{T}\big(\hat{H}_T^{(0)} - \big(\hat{A}^{(j+1)}\hat{H}_T'^{(1)} + \hat{H}_T^{(1)}\hat{A}'^{(j+1)}\big) \\ &\quad + \hat{A}^{(j+1)}\hat{H}_T^{(2)}\hat{A}'^{(j+1)}\big) \\ \widehat{DD'}^{(j+1)} &= \frac{1}{T+1}\left(\sum_{l=0}^{T} y_l y_l' - \big(\hat{J}_T'\hat{C}'^{(j+1)}\right. \\ &\quad \left. + \hat{C}^{(j+1)}\hat{J}_T\big) + \hat{C}^{(j+1)}\hat{H}_T^{(0)}\hat{C}'^{(j+1)}\right) \end{aligned} \tag{8}$$

where $\hat{H}_T^{(M)} \in R^{m\times m}$ for $M = 0, 1, 2$, and $\hat{J}_T \in R^{m\times d}$ are defined, respectively, as

$$\hat{H}_T^{(M)} \triangleq \mathbf{E}_{\hat{\theta}_j}\{H_T^{(M)} \mid \mathcal{Y}_T\}, \qquad \hat{J}_T \triangleq \mathbf{E}_{\hat{\theta}_j}\{J_T \mid \mathcal{Y}_T\}. \tag{9}$$

E. Summary of Main Results

The rest of this paper focuses on deriving recursive finite-dimensional filters for computing $\hat{H}_k^{(M)}$, $M = 0, 1, 2$, and $\hat{J}_k$ at time $k = 1, 2, \cdots$. We assume the general time-varying signal model (1), (2). For convenience, in the sequel, we write $\mathbf{E}_{\hat{\theta}_j}\{\cdot\}$ as $\mathbf{E}\{\cdot\}$, i.e., omit the subscript $\hat{\theta}_j$.

The final form of finite-dimensional filters are summarized in Theorem 7.4, which is the main result of the paper. The theorem holds even if the assumption that B_k is invertible is relaxed as long as the system (1), (2) is uniformly completely controllable (i.e., Definition 7.1 holds).

For the time-invariant state-space model (3), (4), the filter-based EM algorithm for computing the MLE $\theta = (A, B^2, C, DD')$ can be summarized as follows: Choose an initial parameter estimate $\hat{\theta}_0$. At each EM iteration j, $j = 1, 2, \cdots$, compute the estimate $\hat{\theta}_j$, according to (8). The elements of the matrices $\hat{H}_T^{(M)}$, $M = 0, 1, 2$ and $\hat{J}_T$ in (8) are computed as follows (see Theorem 5.4):

$$\begin{aligned} \mathbf{E}\{H_T^{ip(M)} \mid \mathcal{Y}_T\} &= a_T^{ip(M)} + b_T^{ip(M)'}\mu_T \\ &\quad + \mathrm{Tr}\big[d_T^{ip(M)} R_T\big] + \mu_T' d_T^{ip(M)} \mu_T \\ \mathbf{E}\{J_T^{in} \mid \mathcal{Y}_T\} &= \bar{a}_T^{in} + \bar{b}_T^{in'}\mu_T, \\ i = 1, \cdots, m, &\qquad p = 1, \cdots, m, \qquad n = 1 \cdots, d. \end{aligned}$$

The terms μ_T and R_T above are, respectively, the conditional mean and covariance of the state x_T given $\mathcal{Y}_T$. These are computed recursively for $k = 1, 2, \cdots, T$ using the well-known Kalman filter (Theorem 5.1). More importantly, as shown in Theorem 7.4, the terms $a_k^{ip(M)}$, $b_k^{ip(M)}$, etc., in the above equation are sufficient statistics of finite-dimensional filters for computing $\mathbf{E}\{H_k^{ip(M)} \mid \mathcal{Y}_k\}$ and $\mathbf{E}\{J_k^{ip(M)} \mid \mathcal{Y}_k\}$. They are recursively computed for $k = 1, 2, \cdots, T$ according to Theorem 7.4.

The above method for computing the E-step only uses filters—thus we have a filter-based EM algorithm. Another example of a filter-based EM algorithm, for an errors-in-variables model (77), (78), is given by (80) in Section VIII.

III. Measure Change Construction and Dynamics

The aim of this section is to introduce a measure transformation that simplifies the derivation of the filters.

We shall adapt the techniques in [11] and show how the dynamics (1) and (2) can be modeled starting with an initial reference probability measure $\bar{P}$.

Suppose on a probability space $(\Omega, \mathcal{F}, \bar{P})$ we are given two sequences of independent, identically distributed random variables $x_k \in R^m$, $y_k \in R^d$. Under the probability measure $\bar{P}$, the x_k are a sequence of independent m-dimensional $N(0, I_m)$ random variables, and the y_k are a sequence of independent d-dimensional $N(0, I_d)$ random variables. Here, I_m (respectively, I_d) represents the $m\times m$ (respectively, $d\times d$) identity matrix.

For $x \in R^m$ and $y \in R^d$, write

$$\begin{aligned} \psi(x) &= (2\pi)^{-m/2}\exp(-x'x/2) \\ \phi(y) &= (2\pi)^{-d/2}\exp(-y'y/2). \end{aligned} \tag{10}$$

Define the sigma-fields

$$\begin{aligned}\mathcal{G}_k &= \sigma\{x_0, x_1, \cdots, x_k, y_0, y_1, \cdots, y_k\}\\ \mathcal{Y}_k &= \sigma\{y_0, y_1, \cdots, y_k\}.\end{aligned} \tag{11}$$

Thus $\mathcal{G}_k$ is the complete filtration generated by the x and y sequences and $\mathcal{Y}_k$ is the complete filtration generated by the observations.

For any matrix B, let $|B|$ denote its determinant.

Write

$$\lambda_0 = \frac{\phi\big(D_0^{-1}(y_0 - C_0 x_0)\big)}{|D_0|\phi(y_0)}$$

and for $l \geq 1$

$$\lambda_l = \frac{\phi\big(D_l^{-1}(y_l - C_l x_l)\big)}{|D_l|\phi(y_l)} \frac{\psi\big(B_l^{-1}(x_l - A_l x_{l-1})\big)}{|B_l|\psi(x_l)}. \tag{12}$$

For $k \geq 0$ set

$$\Lambda_k = \prod_{l=0}^{k} \lambda_l.$$

A new probability measure P can be defined on $(\Omega, \vee_k \mathcal{G}_k)$ by setting the $\mathcal{G}_k$ restriction of the Radon–Nikodym derivative of P with respect to $\bar{P}$

$$\left.\frac{dP}{d\bar{P}}\right|_{\mathcal{G}_k} = \Lambda_k.$$

Definition 3.1: For $l = 0, 1, \cdots$, define

$$v_l = D_l^{-1}(y_l - C_l x_l).$$

For $l = 1, 2, \cdots$, define

$$w_l = B_l^{-1}(x_l - A_l x_{l-1}). \tag{13}$$

Lemma 3.2: Under the measure P, v_l and w_l are sequences of independent $N(0, I_d)$ and $N(0, I_m)$ random variables, respectively.

The proof appears in the Appendix.

Remark: Note that under the probability measure P, (1) and (2) hold. P represents the "real world" dynamics. However, $\bar{P}$ is a much nicer measure with which to work.

IV. Recursive Estimates

Let e_i, e_j denote the unit column vectors in R^m with 1 in the ith and jth position, respectively. Let e_n be the unit column vector in R^d with 1 in the nth position.

For $i, j \in \{1 \cdots, m\}$ and $n \in \{1, \cdots, d\}$, define the scalar processes

$$\begin{aligned} H_k^{ij(0)} &= \sum_{l=0}^{k} \langle x_l, e_i\rangle\langle x_l, e_j\rangle \quad & H_k^{ij(1)} &= \sum_{l=1}^{k} \langle x_l, e_i\rangle\langle x_{l-1}, e_j\rangle\\ H_k^{ij(2)} &= \sum_{l=1}^{k} \langle x_{l-1}, e_i\rangle\langle x_{l-1}, e_j\rangle \quad & J_k^{in} &= \sum_{l=0}^{k} \langle x_l, e_i\rangle\langle y_l, e_n\rangle \end{aligned} \tag{14}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product. Note that these are merely the elements of the matrices $H_k^{(M)}$, $M = 0, 1, 2$, and J_k.

Our aim is to derive finite-dimensional recursive filters for $H_k^{ij(M)}$, $M = 0, 1, 2$ and J_k^{in}, that is, to compute $\mathbf{E}\{H_k^{ij(M)} \mid \mathcal{Y}_k\}$ and $\mathbf{E}\{J_k^{in} \mid \mathcal{Y}_k\}$ recursively. As shown in Section II, these filtered quantities are required in the filter-based EM algorithm for estimating the parameters.

In order to derive the filters, in this section we derive recursive expressions for the unnormalized densities of x_k, $H_k^{ij(M)}$, $M = 0, 1, 2$ and J_k^{in} under the probability measure $\bar{P}$.

A. Recursive Filtered Densities

Let α_k, $\beta_k^{(0)}$, $\beta_k^{(1)}$, $\beta_k^{(2)}$, and γ_k be the unnormalized (measure valued) densities

$$\begin{aligned}\alpha_k(x) &= \bar{\mathbf{E}}\{\Lambda_k I(x_k \in dx) \mid \mathcal{Y}_k\}\\ \beta_k^{ij(M)}(x) &= \bar{\mathbf{E}}\big\{\Lambda_k H_k^{ij(M)} I(x_k \in dx) \,\big|\, \mathcal{Y}_k\big\}, \quad M = 0, 1, 2\\ \gamma_k^{in}(x) &= \bar{\mathbf{E}}\big\{\Lambda_k J_k^{in} I(x_k \in dx) \,\big|\, \mathcal{Y}_k\big\}.\end{aligned} \tag{15}$$

Then for any measurable "test" function $g: R^m \to R$

$$\begin{aligned}\bar{\mathbf{E}}\{\Lambda_k g(x_k) \mid \mathcal{Y}_k\} &= \int_{R^m} \alpha_k(x) g(x)\, dx\\ \bar{\mathbf{E}}\big\{\Lambda_k H_k^{ij(M)} g(x_k) \,\big|\, \mathcal{Y}_k\big\} &= \int_{R^m} \beta_k^{ij(M)}(x) g(x)\, dx,\\ &\qquad M = 0, 1, 2.\end{aligned} \tag{16}$$

The following theorem gives recursive expressions for the α_k, β_k, and γ_k. The recursions are derived under the measure $\bar{P}$ where $\{x_l\}$ and $\{y_l\}$, $l \in Z^+$ are independent sequences of random variables.

Theorem 4.1: For $k \in Z^+$, the unnormalized densities α_k, $\beta_k^{ij(M)}$, $M = 0, 1, 2$, and γ_k^{in} defined in (15) are given by the following recursions:

$$\alpha_k(x) = \frac{\phi\big(D_k^{-1}(y_k - C_k x)\big)}{|B_k||D_k|\phi(y_k)} \int_{R^m} \alpha_{k-1}(z)\psi\big(B_k^{-1}(x - A_k z)\big)\, dz \tag{17}$$

$$\begin{aligned}\beta_k^{ij(0)}(x) &= \frac{\phi\big(D_k^{-1}(y_k - C_k x)\big)}{|B_k||D_k|\phi(y_k)}\\ &\quad \times \Bigg[\int_{R^m} \beta_{k-1}^{ij(0)}(z)\psi\\ &\quad \times \big(B_k^{-1}(x - A_k z)\big)\, dz + \langle x, e_i\rangle\langle x, e_j\rangle \int_{R^m} \alpha_{k-1}(z)\psi\\ &\quad \times \big(B_k^{-1}(x - A_k z)\big)\, dz\Bigg]\end{aligned} \tag{18}$$

$$\begin{aligned}\beta_k^{ij(1)}(x) &= \frac{\phi\big(D_k^{-1}(y_k - C_k x)\big)}{|B_k||D_k|\phi(y_k)}\\ &\quad \times \Bigg[\int_{R^m} \beta_{k-1}^{ij(0)}(z)\psi\\ &\quad \times \big(B_k^{-1}(x - A_k z)\big)\, dz + \langle x, e_i\rangle \int_{R^m} \langle z, e_j\rangle\alpha_{k-1}(z)\psi\\ &\quad \times \big(B_k^{-1}(x - A_k z)\big)\, dz\Bigg]\end{aligned} \tag{19}$$

$$\begin{aligned}
&\beta_k^{ij(2)}(x)\\
&\quad= \frac{\phi\big(D_k^{-1}(y_k - C_k x)\big)}{|B_k||D_k|\phi(y_k)}\\
&\qquad\times\Bigg[\int_{R^m} \beta_{k-1}^{ij(0)}(z)\psi\\
&\qquad\times\big(B_k^{-1}(x - A_k z)\big)\,dz + \int_{R^m} \langle z, e_i\rangle\langle z, e_j\rangle \alpha_{k-1}(z)\psi\\
&\qquad\times\big(B_k^{-1}(x - A_k z)\big)\,dz\Bigg] \qquad (20)
\end{aligned}$$

$$\begin{aligned}
&\gamma_k^{in}(x)\\
&\quad= \frac{\phi\big(D_k^{-1}(y_k - C_k x)\big)}{|B_k||D_k|\phi(y_k)}\\
&\qquad\Bigg[\int_{R^m} \gamma_{k-1}^{in}(z)\psi\\
&\qquad\times\big(B_k^{-1}(x - A_k z)\big)\,dz + \langle x, e_i\rangle\langle y_k, e_n\rangle \int_{R^m} \alpha_{k-1}(z)\psi\\
&\qquad\times\big(B_k^{-1}(x - A_k z)\big)\,dz\Bigg]. \qquad (21)
\end{aligned}$$

Proof: We prove (18). The proof of (19)–(21) and (17) are very similar and hence omitted.

Since $H_k^{ij(0)} = H_{k-1}^{ij(0)} + \langle x_k, e_i\rangle\langle x_k, e_j\rangle$, using (12) it follows that we have (22), as shown at the bottom of the page, where the second equality follows from the independence of the x_k's and y_k's under $\bar{P}$.

Since g is an arbitrary Borel test function, equating the right-hand side (RHS) of (16) with (22) proves (18). □

Remarks:

1) By virtue of (17), we can rewrite (18) and (21) as

$$\begin{aligned}
\beta_k^{ij(0)}(x) = \frac{\phi\big(D_k^{-1}(y_k - C_k x)\big)}{|B_k||D_k|\phi(y_k)} \int_{R^m} \beta_{k-1}^{ij(0)}(x)\psi\\
\times\big(B_k^{-1}(x - A_k z)\big)\,dz + \langle x, e_i\rangle\langle x, e_j\rangle\alpha_k(x)
\end{aligned} \qquad (23)$$

$$\begin{aligned}
\gamma_k^{in}(x) = \frac{\phi\big(D_k^{-1}(y_k - C_k x)\big)}{|B_k||D_k|\phi(y_k)} \int_{R^m} \gamma_{k-1}^{in}(x)\psi\\
\times\big(B_k^{-1}(x - A_k z)\big)\,dz + \langle x, e_i\rangle\langle y_k, e_n\rangle\alpha_k(x).
\end{aligned} \qquad (24)$$

2) The above theorem does not require v_l and w_l to be Gaussian. The recursions (17), (18), and (21) hold for arbitrary densities ψ and ϕ as long as ϕ is strictly positive. We use the Gaussian assumption to derive the finite-dimensional filters in Section V.

3) *Initial conditions*: Note that at $k = 0$, the following holds for any arbitrary Borel test function $g(x)$:

$$\begin{aligned}
&\bar{\mathbf{E}}\{\Lambda_0 g(x) \mid \mathcal{Y}_0\}\\
&\quad= \bar{\mathbf{E}}\left\{\frac{\phi\big(D_0^{-1}(y_0 - C_0 x)\big)}{|D_0|\phi(y_0)} g(x) \,\middle|\, \mathcal{Y}_0\right\}\\
&\quad= \frac{1}{|D_0|\phi(y_0)} \int_{R^m} \phi\big(D_0^{-1}(y_0 - C_0 x)\big)\psi(x) g(x)\,dx.
\end{aligned} \qquad (25)$$

Equating (15) and (25) yields

$$\alpha_0(x) = \frac{\phi\big(D_0^{-1}(y_0 - C_0 x)\big)}{|D_0|\phi(y_0)}\psi(x). \qquad (26)$$

Similarly the initial conditions for $\beta_k^{ij(M)}$, $M = 0, 1, 2$ and γ_k^{in} are

$$\begin{aligned}
&\beta_0^{ij(0)}(x) = \langle x, e_i\rangle\langle x, e_j\rangle\alpha_0(x) \quad \beta_0^{ij(1)}(x) = 0\\
&\beta_0^{ij(2)}(x) = 0 \quad \gamma_0^{in}(x) = \langle x, e_i\rangle\langle y_0, e_n\rangle\alpha_0(x).
\end{aligned} \qquad (27)$$

V. Finite-Dimensional Filters

In this section finite-dimensional filters are derived for $H_k^{ij(M)}$, $M = 0, 1, 2$ and J_k^{in} defined in (14). In particular, we characterize the densities α_k, $\beta_k^{ij(M)}$ and γ_k^{in} in terms of a finite number of sufficient statistics. Then recursions are derived for these statistics.

$$\begin{aligned}
\bar{\mathbf{E}}\big\{\Lambda_k H_k^{ij(0)} g(x_k) \,\big|\, \mathcal{Y}_k\big\} =& \bar{\mathbf{E}}\left\{\Lambda_{k-1} \frac{\phi\big(D_k^{-1}(y_k - C_k x_k)\big)}{|D_k|\phi(y_k)} \frac{\psi\big(B_k^{-1}(x_k - A_k x_{k-1})\big)}{|B_k|\psi(x_k)} \times H_{k-1}^{ij(0)} g(x_k) \,\middle|\, \mathcal{Y}_k\right\}\\
&+ \bar{\mathbf{E}}\left\{\Lambda_{k-1} \frac{\phi\big(D_k^{-1}(y_k - C_k x_k)\big)}{|D_k|\phi(y_k)} \times \frac{\psi\big(B_k^{-1}(x_k - A_k x_{k-1})\big)}{|B_k|\psi(x_k)} \langle x_k, e_i\rangle\langle x_k, e_j\rangle g(x_k) \,\middle|\, \mathcal{Y}_k\right\}\\
=& \frac{1}{|B_k||D_k|\phi(y_k)} \Bigg[\bar{\mathbf{E}}\bigg\{\Lambda_{k-1} H_{k-1}^{ij(0)} \int_{R^m} \phi\big(D_k^{-1}(y_k - C_k x)\big)\psi\big(B_k^{-1}(x - A_k x_{k-1})\big) g(x) dx \mid \mathcal{Y}_{k-1}\bigg\}\\
&+ \bar{\mathbf{E}}\bigg\{\Lambda_{k-1} \int_{R^m} \phi\big(D_k^{-1}(y_k - C_k x)\big)\,\psi\big(B_k^{-1}(x - A_k x_{k-1})\big)\\
&\times \langle x, e_i\rangle\langle x, e_j\rangle g(x)\,dx \mid \mathcal{Y}_{k-1}\bigg\}\Bigg]\\
=& \frac{1}{|B_k||D_k|\phi(y_k)} \Bigg[\int_{R^m}\int_{R^m} \beta_{k-1}^{ij(0)}(z)\phi\big(D_k^{-1}(y_k - C_k x)\big)\,\psi\big(B_k^{-1}(x - A_k z)\big) g(x)\,dx\,dz\\
&+ \int_{R^m}\int_{R^m} \alpha_{k-1}(z)\phi\big(D_k^{-1}(y_k - C_k x)\big)\,\psi\big(B_k^{-1}(x - A_k z)\big)\langle x, e_i\rangle\langle x, e_j\rangle g(x)\,dx\,dz\Bigg]
\end{aligned} \qquad (22)$$

Define the conditional mean and conditional covariance matrix of x_k, respectively, as $\mu_k = \mathbf{E}\{x_k \mid \mathcal{Y}_k\}$ and $R_k = \mathbf{E}\{(x_k - \mu_k)(x_k - \mu_k)' \mid \mathcal{Y}_k\}$.

The linearity of (1) and (2) implies that $\alpha_k(x)$ is an unnormalized normal density with mean and variance given by the well-known Kalman filter equations.

Theorem 5.1 (Kalman Filter): For $k = 0, 1, \cdots$, α_k is an unnormalized Gaussian density of the form

$$\alpha_k(x) = \bar{\alpha}_k (2\pi)^{-m/2} |R_k|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu_k)' R_k^{-1} (x - \mu_k)\right)$$

where $\bar{\alpha}_k = \int_{R^m} \alpha_k(x)\, dx$. The mean μ_k and covariance R_k are given via the Kalman filter equations

$$\mu_k = R_k B_k^{-2} A_k \sigma_k^{-1} R_{k-1}^{-1} \mu_{k-1} + R_k C_k' (D_k' D_k)^{-1} y_k \quad (28)$$

$$R_k = \left[\left(A_k R_{k-1} A_k' + B_k^2\right)^{-1} + C_k' (D_k D_k')^{-1} C_k\right]^{-1}. \quad (29)$$

Here $\mu_k = \mathbf{E}\{x_k \mid \mathcal{Y}_k\}$ is an m-vector and $R_k = \mathbf{E}\{(x - \mu_k)(x - \mu_k)' \mid \mathcal{Y}_k\}$ is a symmetric $m \times m$ matrix. Also σ_k is a symmetric $m \times m$ matrix defined as

$$\sigma_k = A_k' B_k^{-2} A_k + R_{k-1}^{-1}. \quad (30)$$

Proof: See [11]. □

Due to the presence of the quadratic term $\langle x, e_i \rangle \langle x, e_j \rangle$, the density $\beta_k^{(0)}$ in (23) is not Gaussian. Is it possible to characterize the density $\beta_k^{(0)}$ in terms of a finite number of sufficient statistics? The answer is "yes." As will be proved below, it is possible to express $\beta_k(x)^{(0)}$ as a quadratic expression in x multiplied by $\alpha_k(x)$ for all k. The important conclusion then is that by updating the coefficients of the quadratic expression, together with the Kalman filter above, we have finite-dimensional filters for computing $H_k^{ij(0)}$. A similar result also holds for $H_k^{ij(1)}$, $H_k^{ij(2)}$, and J_k^{in}.

Theorems 5.2 and 5.3 that follow derive finite-dimensional sufficient statistics for the densities $\beta_k^{ij(M)}$, $M = 0, 1, 2$, and γ_k^{in}. To simplify the notation, we define

$$\Sigma_k = B_k^{-2} A_k \sigma_k^{-1}, \qquad S_k = \sigma_{k+1}^{-1} R_k^{-1} \mu_k. \quad (31)$$

Theorem 5.2: At time k, the density $\beta_k^{ij(M)}(x)$ [initialized according to (27)] is completely defined by the five statistics $a_k^{ij(M)}$, $b_k^{ij(M)}$, $d_k^{ij(M)}$, R_k, and μ_k as follows:

$$\beta_k^{ij(M)}(x) = \left[a_k^{ij(M)} + b_k^{ij(M)'} x + x' d_k^{ij(M)} x\right] \alpha_k(x), \qquad M = 0, 1, 2 \quad (32)$$

where $a_k^{ij(M)} \in R$, $b_k^{ij(M)} \in R^m$ and $d_k^{ij(M)} \in R^{m \times m}$ is a symmetric matrix with elements $d_k(p, q)$, $p = 1, \cdots, m$, $q = 1, \cdots, m$.

Furthermore, $a_k^{ij(M)}$, $b_k^{ij(M)}$, and $d_k^{ij(M)}$, $M = 0, 1, 2$ are given by the following recursions:

$$a_{k+1}^{ij(0)} = a_k^{ij(0)} + b_k^{ij(0)'} S_k + \mathrm{Tr}\left[d_k^{ij(0)} \sigma_{k+1}^{-1}\right] + S_k' d_k^{ij(0)} S_k, \qquad a_0^{ij(0)} = 0 \quad (33)$$

$$b_{k+1}^{ij(0)} = \Sigma_{k+1}\left(b_k^{ij(0)} + 2 d_k^{ij(0)} S_k\right) \quad b_0^{ij(0)} = 0_{m \times 1} \quad (34)$$

$$d_{k+1}^{ij(0)} = \Sigma_{k+1} d_k^{ij(0)} \Sigma_{k+1}' + \frac{1}{2}(e_i e_j' + e_j e_i')$$

$$d_0^{ij(0)} = \frac{e_i e_j' + e_j e_i'}{2} \quad (35)$$

$$a_{k+1}^{ij(1)} = a_k^{ij(1)} + b_k^{ij(1)'} S_k + \mathrm{Tr}\left[d_k^{ij(1)} \sigma_{k+1}^{-1}\right] + S_k' d_k^{ij(1)} S_k, \qquad a_0^{ij(1)} = 0 \quad (36)$$

$$b_{k+1}^{ij(1)} = \Sigma_{k+1}\left(b_k^{ij(1)} + 2 d_k^{ij(1)} S_k\right) + e_i e_j' S_k \quad b_0^{ij(1)} = 0_{m \times 1} \quad (37)$$

$$d_{k+1}^{ij(1)} = \Sigma_{k+1} d_k^{ij(1)} \Sigma_{k+1}' + \frac{1}{2}\left[e_i e_j' \Sigma_{k+1}' + \Sigma_{k+1} e_j e_i'\right] \qquad d_0^{ij(1)} = 0_{m \times m} \quad (38)$$

$$a_{k+1}^{ij(2)} = a_k^{ij(2)} + b_k^{ij(2)'} S_k + \mathrm{Tr}\left[d_k^{ij(2)} \sigma_{k+1}^{-1}\right] + S_k'\left(d_k^{ij(2)} + e_i e_j'\right) S_k + \mathrm{Tr}\left[e_i e_j' \sigma_{k+1}^{-1}\right], \quad a_0^{ij(2)} = 0 \quad (39)$$

$$b_{k+1}^{ij(2)} = \Sigma_{k+1}\left(b_k^{ij(2)} + \left(2 d_k^{ij(2)} + 2 e_j e_i'\right) S_k\right) \quad b_0^{ij(2)} = 0_{m \times 1} \quad (40)$$

$$d_{k+1}^{ij(2)} = \Sigma_{k+1}\left(d_k^{ij(2)} + \frac{1}{2}[e_i e_j' + e_i' e_j]\right) \Sigma_{k+1}' \qquad d_0^{ij(2)} = 0_{m \times m} \quad (41)$$

where $\mathrm{Tr}[\cdot]$ denotes the trace of a matrix, σ_k is defined in (30), and μ_k, R_k are obtained from the Kalman filter (28) and (29).

Proof: We only prove the theorem for $M = 0$; the proofs for $M = 1, 2$ are very similar and hence omitted.

We prove (32) by induction.

From (27), at time $k = 0$, $\beta_0^{ij(0)}(x)$ is of the form (32) with $a_0^{ij(0)} = 0$, $b_0^{ij(0)} = 0$, and $d_0^{ij(0)} = (e_i e_j' + e_j e_i')/2$.

For convenience we drop the superscripts $i, j, (0)$ in $a_k^{ij(0)}$, $b_k^{ij(0)}$, $d_k^{ij(0)}$, and $\beta_{k+1}^{ij(0)}$ Assume that (32) holds at time k. Then at time $k+1$, using (32) and the recursion (23) it follows that

$$\beta_{k+1}(x) = \frac{\phi\left(D_{k+1}^{-1}(y_{k+1} - C_{k+1} x)\right)}{|B_{k+1}||D_{k+1}|\phi(y_{k+1})} \int_{R^m} \psi\left(B_{k+1}^{-1}(x - A_{k+1} z)\right) \times (a_k + b_k' z + z' d_k z) \alpha_k(z)\, dz + \langle x, e_i \rangle \langle x, e_j \rangle \alpha_{k+1}(x). \quad (42)$$

Let us concentrate on the first term on the RHS which we shall denote as I_1

$$\begin{aligned} I_1 &= K(x) \int_{R^m} \exp\left[-\frac{1}{2}\{(x - A_{k+1} z)' B_{k+1}^{-2} (x - A_{k+1} z) + (z - \mu_k)' R_k^{-1} (z - \mu_k)\}\right] \\ &\quad \times (a_k + b_k' z + z' d_k z)\, dz \\ &= K_1(x) \int_{R^m} \exp\left[-\frac{1}{2}(z' \sigma_{k+1} z - \delta_{k+1}' z)\right] \\ &\quad \times (a_k + b_k' z + z' d_k z)\, dz \end{aligned} \quad (43)$$

where σ_k is defined in (30)

$$\begin{aligned}\delta_{k+1} &= 2\big(x'B_{k+1}^{-2}A_{k+1} + \mu_k' R_k^{-1}\big)' \\ K(x) &= \frac{\phi\big(D_{k+1}^{-1}(y_{k+1} - C_{k+1}x)\big)}{|B_{k+1}||D_{k+1}|\phi(y_{k+1})}(2\pi)^{-m}|B_{k+1}|^{-1} \\ &\quad \times |R_k|^{-1/2}\bar{\alpha}_k \\ \bar{\alpha}_k &= \int_{R^m} \alpha_k(z)\, dz \\ K_1(x) &= K(x)\exp\left[-\frac{1}{2}\big(x'B_{k+1}^{-2}x + \mu_k' R_k^{-1}\mu_k\big)\right]. \end{aligned} \tag{44}$$

Completing the "square" in the exponential term in (43) yields

$$\begin{aligned} I_1 = K_1(x)&\exp\left[-\frac{1}{2}\left(-\frac{\delta_{k+1}'\sigma_{k+1}^{-1}\delta_{k+1}}{4}\right)\right] \\ &\times \int_{R^m} \exp\left[-\frac{1}{2}\left(z - \frac{\sigma_{k+1}^{-1}\delta_{k+1}}{2}\right)'\sigma_{k+1}\right. \\ &\qquad \left.\times\left(z - \frac{\sigma_{k+1}^{-1}\delta_{k+1}}{2}\right)\right](a_k + b_k' z + z' d_k z)\, dz. \end{aligned} \tag{45}$$

Now consider the integral in (45)

$$\begin{aligned} \int_{R^m} (a_k + b_k' z + z' d_k z)&\exp\left[-\frac{1}{2}\left(z - \frac{\sigma_{k+1}^{-1}\delta_{k+1}}{2}\right)'\sigma_{k+1}\right. \\ &\quad \left.\times\left(z - \frac{\sigma_{k+1}^{-1}\delta_{k+1}}{2}\right)\right] dz \\ &= (2\pi)^{m/2}|\sigma_{k+1}|^{-1/2}(a_k + b_k'\mathbf{E}\{z\} + \mathbf{E}\{z' d_k z\}) \end{aligned} \tag{46}$$

since the $\exp(\cdot)$ term is an unnormalized Gaussian density in z with normalization constant $(2\pi)^{m/2}|\sigma_{k+1}|^{-1/2}$. (Here $\mathbf{E}\{z\}$ denotes the expected value of the Gaussian random variable z.) So

$$\mathbf{E}\{z\} = \frac{\sigma_{k+1}^{-1}\delta_{k+1}}{2} \tag{47}$$

$$\begin{aligned} &\mathbf{E}\{z' d_k z\} \\ &\quad = \mathbf{E}\{(z - \mathbf{E}\{z\})' d_k (z - \mathbf{E}\{z\})\} + \mathbf{E}\{z'\}\, d_k \mathbf{E}\{z\} \\ &\quad = \mathrm{Tr}\big[d_k \sigma_{k+1}^{-1}\big] + \frac{1}{4}\big(\sigma_{k+1}^{-1}\delta_{k+1}\big)' d_k \big(\sigma_{k+1}^{-1}\delta_{k+1}\big). \end{aligned} \tag{48}$$

Therefore from (45)–(48) and (42) it follows that

$$\begin{aligned} \beta_{k+1}(x) = \alpha_{k+1}(x)&\left[a_k + \frac{1}{2}b_k'\sigma_{k+1}^{-1}\delta_{k+1} + \mathrm{Tr}\big[d_k\sigma_{k+1}^{-1}\big]\right. \\ &\quad \left. + \frac{1}{4}\delta_{k+1}'\sigma_{k+1}^{-1}d_k\sigma_{k+1}^{-1}\delta_{k+1} + x'e_ie_j'x\right]. \end{aligned} \tag{49}$$

Substituting for δ_{k+1} (which is affine in x) in (49) yields

$$\beta_{k+1}(x) = (a_{k+1} + b_{k+1}'x + x'd_{k+1}x)\alpha_{k+1}(x)$$

where a_{k+1}, b_{k+1}, and d_{k+1} are given by (33)–(35). □

The proof of the following theorem is very similar and hence omitted.

Theorem 5.3: The density $\gamma_k^{in}(x)$ is completely determined by the four statistics $\bar{a}_k^{in}$, $\bar{b}_k^{in}$, $\mu_k \in R^m$, and $R_k \in R^{m\times m}$ as follows:

$$\begin{aligned} \gamma_k^{in}(x) &= \big[\bar{a}_k^{in} + \bar{b}_k^{in'}x\big]\alpha_k(x) \\ \gamma_0^{in}(x) &= \langle x, e_i\rangle\langle y_0, e_n\rangle\alpha_0(x) \end{aligned} \tag{50}$$

where $\bar{a}_k^{in} \in R$, $\bar{b}_k^{in} \in R^m$ are given by the following recursions:

$$\bar{a}_{k+1}^{in} = \bar{a}_k^{in} + \bar{b}_k^{in'}S_k, \qquad \bar{a}_0^{in} = 0 \tag{51}$$

$$\bar{b}_{k+1}^{in} = \Sigma_{k+1}\bar{b}_k^{in} + e_i\langle y_{k+1}, e_n\rangle, \qquad \bar{b}_0^{in} = e_i\langle y_0, e_n\rangle. \tag{52}$$

where Σ_k and S_k are defined in (31).

Having characterized the densities $\beta_k^{ij(M)}(x)$, $M = 0, 1, 2$, and $\gamma_k^{in}(x)$ by their finite sufficient statistics, we now derive finite-dimensional filters for $H_k^{ij(M)}$ and J_k^{in}.

Theorem 5.4: Finite-dimensional filters for $H_k^{ij(M)}$, $M = 0, 1, 2$, and J_k^{in} are given by

$$\begin{aligned} \mathbf{E}\{H_k^{ij(M)} \,|\, \mathcal{Y}_k\} &= a_k^{ij(M)} + b_k^{ij(M)'}\mu_k + \mathrm{Tr}\big[d_k^{ij(M)}R_k\big] \\ &\qquad + \mu_k' d_k^{ij(M)}\mu_k \end{aligned} \tag{53}$$

$$\mathbf{E}\{J_k^{in} \,|\, \mathcal{Y}_k\} = \bar{a}_k^{in} + \bar{b}_k^{in'}\mu_k. \tag{54}$$

Proof: Using the abstract Bayes rule (81) it follows that

$$\mathbf{E}\{H_k^{ij(M)} \,|\, \mathcal{Y}_k\} = \frac{\bar{\mathbf{E}}\{\Lambda_k H_k^{ij(M)} \,|\, \mathcal{Y}_k\}}{\bar{\mathbf{E}}\{\Lambda_k \mid \mathcal{Y}_k\}} = \frac{\int_{R^m}\beta_k^{ij(M)}(x)\, dx}{K} \tag{55}$$

where the constant $K = \int_{R^m}\alpha_k(x)\, dx$. But since $\alpha_k(x)$ is an unnormalized density, from (32)

$$\begin{aligned} &\int_{R^m}\beta_k^{ij(M)}(x)\, dx \\ &\quad = K\mathbf{E}\{a_k^{ij(M)} + b_k^{ij(M)'}x + x'd_k^{ij(M)}x\} \\ &\quad = K\big[a_k^{ij(M)} + b_k^{ij(M)'}\mu_k + \mathrm{Tr}\big[d_k^{ij(M)}R_k\big] + \mu_k' d_k^{ij(M)}\mu_k\big]. \end{aligned} \tag{56}$$

Substituting in (55) proves the theorem.

The proof of (54) is similar and hence omitted. □

VI. General Filter for Higher Order Moments

Theorem 5.4 gives finite-dimensional filters for the time sum of the states J_k^{in} and time sum of the square of the states H_k^{ij}. In this section we show that *finite-dimensional filters exist for the time sum of any arbitrary integral power of the states.*

Assumption 6.1: For notational simplicity, in this section we assume that the state and observation processes are scalar valued, i.e., $m = d = 1$ in (1) and (2).

Let H_k be the time sum of the pth power of the state[1]

$$H_k \triangleq \sum_{l=0}^{k} x_l^p, \qquad p \in Z^+. \tag{57}$$

Our aim is to derive a finite-dimensional filter for H_k.

[1] These new definitions for H_k in (57) and $\beta_k(x)$ in (58) are only used in this section.

Define the unnormalized density $\beta_k(x) = \bar{\mathbf{E}}\{\Lambda_k H_k I(x_k \in dx) \mid \mathcal{Y}_k\}$.

Our first step is to obtain a recursion for $\beta_k(x)$.

By using a proof very similar to Theorem 4.1, we can show

$$\beta_k(x) = \frac{\phi(D_k^{-1}(y_k - C_k x))}{|B_k||D_k|\phi(y_k)} \int_R \beta_{k-1}(z) \times \psi\big(B_k^{-1}(x - A_k z)\big)\, dz + x^p \alpha_k(x). \tag{58}$$

Our task now is to characterize $\beta_k(x)$ in terms of finite sufficient statistics. Recall that for $p = 0$, the Kalman filter state and covariance are sufficient statistics as shown in Theorem 5.1. Also for $p = 1$ and 2, Theorems 5.3 and 5.2 give finite-dimensional sufficient statistics. We now show that β_k can be characterized in terms of finite-dimensional statistics for any $p \in Z^+$.

Theorem 6.2: At time k, the density $\beta_k(x)$ in (58) is completely defined by the $p+3$ statistics $a_k(0), a_k(1), \cdots, a_k(p)$, R_k and μ_k as follows:

$$\beta_k(x) = \left[\sum_{i=0}^{p} a_k(i) x^i\right] \alpha_k(x) \tag{59}$$

where

$$a_{k+1}(n) = \sum_{i=n}^{p}\sum_{j=n}^{i} a_k(i)\eta_{ij}\binom{j}{n}\big(R_k^{-1}\mu_k\big)^{j-n}\big(A_{k+1}B_{k+1}^{-2}\big)^n \qquad 0 \le n < p$$

$$a_{k+1}(p) = 1 + a_k(p)\eta_{pp}\big(A_{k+1}B_{k+1}^{-2}\big)^p \tag{60}$$

and (61), as shown at the bottom of the page.

Proof: As in Theorem 5.2, we give an inductive proof. At $k = 0$, $\beta_k(x) = x^p \alpha_0(x)$ and thus satisfies (59).

Assume that (59) holds at time k. Then at time $k+1$, using similar arguments to Theorem 5.2, it follows that

$$\beta_{k+1}(x) = \frac{\phi(D_{k+1}^{-1}(y_{k+1} - C_{k+1}x))}{|B_{k+1}||D_{k+1}|\phi(y_{k+1})} \int_R \psi\big(B_{k+1}^{-1}(x - A_{k+1}z)\big) \times \left(\sum_{i=0}^{p} a_k(i) z^i\right)\alpha_k(z)\, dz + x^p \alpha_k(x). \tag{62}$$

The first term on the RHS of the above equation is

$$I_1 = K_1(x) \exp\left[-\frac{1}{2}\left(-\frac{\delta'_{k+1}\sigma_{k+1}^{-1}\delta_{k+1}}{4}\right)\right] \times \int_{R^m} \exp\left[-\frac{1}{2}\left(z - \frac{\sigma_{k+1}^{-1}\delta_{k+1}}{2}\right)^2 \sigma_{k+1}\right] \times \left(\sum_{i=0}^{p} a_k(i) z^i\right) dz. \tag{63}$$

The integral in the above equation is

$$(2\pi)^{1/2}|\sigma_{k+1}|^{-1/2}\mathbf{E}\left\{\sum_{i=0}^{p} a_k(i) z^i\right\} = (2\pi)^{1/2}|\sigma_{k+1}|^{-1/2}\sum_{i=0}^{p} a_k(i) \times \sum_{j=0}^{i}\binom{i}{j}\mathbf{E}\{(z - \mathbf{E}\{z\})^{i-j}\}(\mathbf{E}\{z\})^j. \tag{64}$$

Now recall from (47) that $\mathbf{E}\{z\}$ is affine in x

$$\mathbf{E}\{z\} = \sigma_{k+1}^{-1}\big[R_k^{-1}\mu_k + A_{k+1}B_{k+1}^{-2}x\big]. \tag{65}$$

Also $\mathbf{E}\{(z - \mathbf{E}\{z\})^2\}$ is independent of x. Indeed, ([12, p. 111])

$$\mathbf{E}\{(z - \mathbf{E}\{z\})^{i-j}\} = \begin{cases} 0, & \text{if } i-j \text{ is odd, } i > j \\ 1\cdot 3\cdots(i-j-1)\sigma_{k+1}^{-1}, & \text{if } i-j \text{ is even, } i > j \\ 1, & \text{if } i = j. \end{cases} \tag{66}$$

Thus

$$\beta_{k+1}(x) = \alpha_{k+1}(x)\left[\sum_{i=0}^{p}\sum_{j=0}^{p}\sum_{n=0}^{j} a_k(i)\eta_{ij}\binom{j}{n}\big(R_k^{-1}\mu_k\big)^{j-n} \times \big(A_{k+1}B_{k+1}^{-2}\big)^n x^n + x^p\right]$$
$$= \alpha_{k+1}(x)\left[\sum_{n=0}^{p}\sum_{i=n}^{p}\sum_{j=n}^{i} a_k(i)\eta_{ij}\binom{j}{n}\big(R_k^{-1}\mu_k\big)^{j-n} \times \big(A_{k+1}B_{k+1}^{-2}\big)^n x^n + x^p\right]. \tag{67}$$

Equation (67) is of the form (59) with $a_{k+1}(i)$, $i = 0, \cdots, p$ given by (60). □

VII. Singular State Noise

The filters derived in Theorems 5.1, 5.2, and 5.3 have one major problem: They require B_k to be invertible. In practice (e.g., see Section VIII), B_k is often not invertible.

In this section, we will use a simple transformation that expresses the filters in the terms of the inverse of the predicted Kalman covariance matrix. This inverse exists even if B_k is singular as long as a certain uniform controllability condition holds. Both the uniform controllability condition and the transformation we use are well known in the Kalman filter literature [15, Ch. 7].

$$\eta_{ij} = \begin{cases} \binom{i}{j} 1\cdot 3\cdots(i-j-1)\sigma_{k+1}^{-(j+1)}, & \text{if } i-j \text{ is even, } i > j \\ 0, & \text{if } i-j \text{ is odd, } i > j \\ \sigma_{k+1}^{-j}, & \text{if } i = j \end{cases} \tag{61}$$

First define the Kalman predicted state estimate $\mu_{k|k-1} \triangleq \mathbf{E}\{x_k \mid \mathcal{Y}_{k-1}\}$ and the predicted state covariance $R_{k|k-1} \triangleq \mathbf{E}\{(x_k - \mu_{k|k-1})(x_k - \mu_{k|k-1})' \mid \mathcal{Y}_{k-1}\}$. It is straightforward to show that

$$R_{k|k-1} = B_k^2 + A_k R_{k-1} A_k' \tag{68}$$

where R_{k-1} denotes the filtered state covariance at time $k-1$ [see (29)].

Our first step is to provide a sufficient condition for $R_{k|k-1}$ to be nonsingular.

Definition 7.1 [15, Ch. 7]: The state-space model (1), (2) is said to be uniformly completely controllable if there exist a positive integer N_1 and positive constants α, β such that

$$\alpha I \leq \mathcal{C}(k, k-N_1) \leq \beta I \quad \text{for all } k \geq N_1. \tag{69}$$

Here

$$\mathcal{C}(k, k-N_1) \triangleq \sum_{l=k-N_1}^{k} \phi(k, l+1) B_l B_l' \phi'(k, l+1) \tag{70}$$

$$\phi(k_2, k_1) = \begin{cases} A_{k_2} A_{k_2-1} \cdots A_{k_1+1}, & \text{if } k_2 > k_1 \\ I, & \text{if } k_2 = k_1 \end{cases}. \tag{71}$$

Lemma 7.2: If the dynamical system (1), (2) is uniformly completely controllable and $R_0 \geq 0$, then R_k and $R_{k|k-1}$ are positive definite matrices (and hence nonsingular) for all $k \geq N_1$.

Proof: See [15, p. 238, Lemma 7.3]. □

Our aim now is to re-express the filters in Section V in terms of $R_{k|k-1}$. The following lemma will be used in the sequel.

Lemma 7.3: Assume $R_{k|k-1}^{-1}$ exists. Then with σ_k and Σ_k defined in (30) and (31), respectively

$$\sigma_k^{-1} = R_{k-1} - R_{k-1} A_k' R_{k|k-1}^{-1} A_k R_{k-1} \tag{72}$$

$$\Sigma_{k+1} = R_{k+1|k}^{-1} A_{k+1} R_k. \tag{73}$$

Furthermore, the Kalman filter (28), (29) can be expressed in "standard" form as

$$\begin{aligned} \mu_k &= A_k \mu_{k-1} + R_{k|k-1} C_k' \left[C_k R_{k|k-1} C_k' + D_k D_k' \right]^{-1} \\ &\quad \times (y_k - C_k A_k \mu_{k-1}) \\ R_k &= R_{k|k-1} - R_{k|k-1} C_k' \left(C_k R_{k|k-1} C_k' + D_k D_k' \right)^{-1} \\ &\quad \times C_k R_{k|k-1} \\ R_{k|k-1} &= B_k^2 + A_k R_{k-1} A_k'. \end{aligned} \tag{74}$$

Proof: Straightforward use of the matrix inversion lemma on (30) yields

$$\sigma_k^{-1} = R_{k-1} - R_{k-1} A_k' \left(B_k^2 + A_k R_{k-1} A_k' \right)^{-1} A_k R_{k-1}. \tag{75}$$

Substituting (68) in (75) proves (72).

To prove (73), first note that

$$\begin{aligned} \Sigma_{k+1} &\triangleq B_{k+1}^{-2} A_{k+1} \sigma_{k+1}^{-1} \\ &= B_{k+1}^{-2} A_{k+1} R_k - B_{k+1}^{-2} A_{k+1} R_k A_{k+1}' R_{k+1|k}^{-1} A_{k+1} R_k \\ &= B_{k+1}^{-2} A_{k+1} R_k - B_{k+1}^{-2} \left(R_{k+1|k} - B_{k+1}^2 \right) R_{k+1|k}^{-1} A_{k+1} R_k \end{aligned}$$

because $A_{k+1} R_k A_{k+1}' = R_{k+1|k} - B_{k+1}^2$ from (68). So

$$\begin{aligned} \Sigma_{k+1} &= B_{k+1}^{-2} A_{k+1} R_k - B_{k+1}^{-2} A_{k+1} R_k + R_{k+1|k}^{-1} A_{k+1} R_k \\ &= R_{k+1|k}^{-1} A_{k+1} R_k. \end{aligned}$$

To prove (74), consider the Kalman filter equations (28) and (29). Using Lemma 7.3 on (29) gives

$$R_k = \left(R_{k|k-1}^{-1} + C_k' (D_k D_k')^{-1} C_k \right)^{-1}. \tag{76}$$

Using the matrix inversion lemma on (76) and applying (73) to the first term on the RHS of (28) yields the "standard" Kalman filter equations. □

Applying the above lemma to the filters derived in Section V, we now express them in terms of $R_{k|k-1}$ instead of B_k. As shown below, the advantage of doing so is that B_k no longer needs to be invertible, as long as the uniformly controllability condition in Definition 7.1 holds.

The following theorem gives the finite-dimensional filters for $H_k^{ij(M)}$, $M = 0, 1, 2$ and J_k^{in} defined in (14) and is the main result of this paper.

Theorem 7.4: Consider the linear dynamical system (1) and (2) with B_k not necessarily invertible. Assume that the system is uniformly completely controllable, i.e., (69) holds. Then at time k, with σ_k^{-1} given by (72) and Σ_k defined in (73), the following hold.

1) The density $\alpha_k(x)$ [defined in (15)] is an unnormalized Gaussian density with mean $\mu_k \in R^m$ and covariance $R_k \in R^{m \times m}$. These are recursively computed via the standard Kalman filter equations (74).

2) The density $\beta_k^{ij(M)}(x)$ [defined in (15) and initialized according to (27)] is completely defined by the five statistics $a_k^{ij(M)}$, $b_k^{ij(M)}$, $d_k^{ij(M)}$, R_k and μ_k as follows:

$$\beta_k^{ij(M)}(x) = \left[a_k^{ij(M)} + b_k^{ij(M)'} x + x' d_k^{ij(M)} x \right] \alpha_k(x), \quad M = 0, 1, 2$$

where $a_k^{ij(M)} \in R$, $b_k^{ij(M)} \in R^m$, $d_k^{ij(M)} \in R^{m \times m}$ is a symmetric matrix with elements $d_k(p, q)$, $p = 1, \cdots, m$, $q = 1, \cdots, m$. These statistics are recursively computed by (33)–(41).

3) The density $\gamma_k^{in}(x)$ [defined in (15)] is completely determined by the four statistics $\bar{a}_k^{in}$, $\bar{b}_k^{in}$ as follows:

$$\begin{aligned} \gamma_k^{in}(x) &= \left[\bar{a}_k^{in} + \bar{b}_k^{in'} x \right] \alpha_k(x) \\ \gamma_0^{in}(x) &= \langle x, e_i \rangle \langle y_0, e_n \rangle \alpha_0(x) \end{aligned}$$

where $\bar{a}_k^{in} \in R$, $\bar{b}_k^{in} \in R^m$. These statistics are recursively computed via (51), (52).

Finally, finite-dimensional filters for $H_k^{ij(M)}$ and J_k^{in} [defined in (14)] in terms of the above statistics are given by (53) and (54).

Proof: It only remains to show that subject to the uniform complete controllability condition (69), the filtering equations

(33)–(41) and (51), (52) in Theorem 7.4 hold even if the matrices B_{k+1} are singular. The proof of this is as follows: If B_{k+1} is singular, then do the following.

1) Add $\epsilon \times N(0,1)$ noise to each component of x_{k+1}. This is done by replacing B_{k+1} in (1) with the nonsingular matrix $B^{\epsilon}_{k+1} = B_{k+1} + \epsilon I_m$ where $\epsilon \in R$. Denote the resulting state process as x^{ϵ}_{k+1}.
2) Define $R^{\epsilon}_{k+1|k}$ as in (68) with B_{k+1} replaced by B^{ϵ}_{k+1}. Express the filters in terms of $R^{\epsilon}_{k+1|k}$ as in Theorem 7.4.
3) As $\epsilon \to 0$, $R^{\epsilon}_{k+1|k} \to R_{k+1|k}$.
4) Then using the bounded conditional convergence theorem ([16, p. 214]), the conditional estimates of x^{ϵ}_k, $x^{\epsilon}_k x^{\epsilon}_k{}'$, $H^{ij(0)}_k(x^{\epsilon})$, and $J^{in}_k(x^{\epsilon})$ converge to the conditional estimates of x_k, $x_k x'_k$, $H^{ij(0)}_k(x)$, and $J^{in}_k(x)$, respectively. □

VIII. Example: MLE of Errors-in-Variables Time Series

We now illustrate the use of the filtered EM algorithm to estimate the parameters of the errors-in-variables time series example considered in [2] and [4].

Consider the scalar valued AR(p) process s_k, $k \in Z^+$, defined as

$$s_k = \sum_{i=1}^{p} a_i s_{k-i} + \nu_k, \qquad \nu_k \sim N(0, \sigma^2_\nu) \tag{77}$$

where ν_k is a white Gaussian process. Assume that s_k is observed indirectly via the scalar process

$$y_k = s_k + \epsilon_k, \qquad \epsilon_k \sim N(0, \sigma^2_\epsilon) \tag{78}$$

where ϵ_k is a white Gaussian process independent of ν_k. The aim is to compute the MLE of the parameter vector $\theta = (a_1, \cdots, a_p, \sigma^2_\nu, \sigma^2_\epsilon)$ using the filter-based EM algorithm.

We first re-express (77) and (78) in state-space form (3), (4) with $d = 1$, $m = p + 1$, $a = [a_1 \cdots a_p]$

$$x_k = \begin{bmatrix} s_k \\ \vdots \\ s_{k-p} \end{bmatrix}, \qquad A = \begin{bmatrix} a & 0 \\ I_{p\times p} & 0_{p\times 1} \end{bmatrix}$$

$$B = \begin{bmatrix} \sigma_\nu & 0_{1\times p} \\ 0_{p\times 1} & 0_{p\times p} \end{bmatrix}, \qquad C = [1 \quad 0_{1\times p}], \qquad D = \sigma_\epsilon.$$

Using a similar procedure to (5), it can be shown [2], [4], [5] that the E-step yields

$$\begin{aligned} Q(\theta, \hat{\theta}_j) &= -\frac{T}{2}\log\sigma_\nu - \frac{1}{2\sigma^2_\nu}\sum_{l=1}^{T}\mathbf{E}_{\hat{\theta}_j}\left\{\left(s_l - \sum_{i=1}^{p} a_i s_{l-i}\right)^2 \middle| \mathcal{Y}_T\right\} \\ &\quad - \frac{T+1}{2}\log\sigma_\epsilon - \frac{1}{2\sigma^2_\epsilon}\sum_{l=0}^{T}\mathbf{E}_{\hat{\theta}_j}\{(y_l - s_l)^2 \mid \mathcal{Y}_T\} \\ &\quad + \mathbf{E}_{\hat{\theta}_j}\{R(\hat{\theta}_j) \mid \mathcal{Y}_T\} \end{aligned} \tag{79}$$

where $R(\hat{\theta}_j)$ does not involve θ.

The M-step yields [2], [4], [5]

$$\begin{aligned} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} &= \begin{bmatrix} \hat{H}^{2,2(0)}_T & \hat{H}^{2,3(0)}_T & \cdots & \hat{H}^{2,p+1(0)}_T \\ & \hat{H}^{3,3(0)}_T & \cdots & \hat{H}^{3,p+1(0)}_T \\ & \text{symmetric} & \ddots & \vdots \\ & & & \hat{H}^{p+1,p+1(0)}_T \end{bmatrix}^{-1} \\ &\quad \times \begin{bmatrix} \hat{H}^{1,2(0)}_T \\ \hat{H}^{1,3(0)}_T \\ \vdots \\ \hat{H}^{1,p+1(0)}_T \end{bmatrix} \\ \sigma^2_\nu &= \frac{1}{T}\left(\hat{H}^{11(0)}_T - \sum_{i=1}^{p}\hat{H}^{i+1,1(0)}_T a_i - \sum_{i=1}^{p}\hat{H}^{1,i+1(0)}_T a_i \right. \\ &\quad \left. + \sum_{i=1}^{p}\sum_{j=1}^{p}\hat{H}^{ij(0)}_T a_i a_j\right) \\ \sigma^2_\epsilon &= \frac{1}{T+1}\left(\sum_{l=0}^{T} y^2_l + \hat{H}^{11(0)}_T y_l - 2\hat{J}^{11}_T\right) \end{aligned} \tag{80}$$

where $\hat{H}^{(0)}_T$ and $\hat{J}_T$ are defined in (9) and computed using (53), (54) together with the finite-dimensional recursive filters in Theorem 7.4.

Proving that the EM algorithm converges to a stationary point on the likelihood surface requires verification of the conditions stated on Section II-A. If σ^2_ν and σ^2_ϵ are assumed known, these conditions are straightforward to verify. Otherwise, it is necessary to ensure that these variances are strictly positive (kept away from zero); see [2] or [14] for details.

Strong consistency of the MLE under the conditions that θ lies in a compact set and that the roots of $z^p - a_1 z^{p-1} - \cdots - a_p = 0$ lie inside the unit circle (i.e., stationarity) is proved in [10, Ch. 7].

Similar parameterized models are used in [1] and [7] and can be estimated via the filter-based EM algorithm presented in this paper.

IX. Parallel Implementation of Filters

In this section we discuss the computational complexity of the new filters and the resulting filter-based EM algorithm. In particular, we describe why the filter-based algorithm is suitable for parallel implementation and propose a systolic processor implementation of the algorithm.

A. Sequential Complexity

We evaluate the computational cost and memory requirements of the filter-based algorithm and compare them with the standard smoother-based EM algorithm.

Computational Cost: The filter-based E-step requires computation at each time k of $\mathbf{E}\{H^{ij(M)}_k \mid \mathcal{Y}_k\}$, $M = 0, 1, 2$, and $\mathbf{E}\{J^{in}_k \mid \mathcal{Y}_k\}$ for all pairs (i, j).

- $a^{ij(M)}_{k+1}$: Consider the RHS of the update (33). The following are the computational cost for each (i, j) pair at each time-instant k.

  Second term: $O(m)$ multiplications (inner product of two m-vectors).

Third term: $O(m^2)$ multiplications (matrix vector multiplication).

Fourth term: $O(m^2)$ multiplications (matrix vector multiplication).

Since there are m^2 (i, j) pairs, the total complexity at each time instant is $O(m^4)$.

- Similarly the total complexity for evaluating $b_{k+1}^{ij(M)}$ for all m^2 (i, j) pairs is $O(m^4)$ multiplications.
- Evaluating $d_{k+1}^{ij(M)}$ for each (i, j) pair requires multiplication of a fixed number of $m \times m$ matrices. This involves $O(m^3)$ complexity. So the total complexity for all m^2 (i, j) pairs is $O(m^5)$ at each time instant.

In comparison, the Kalman smoother-based E-step in [2] and [4] requires $O(m^3)$ complexity at each time instant to compute $\mathbf{E}\{H_T^{ij(M)} \mid \mathcal{Y}_T\}$, $M = 0, 1, 2$, and $\mathbf{E}\{J_T^{in} \mid \mathcal{Y}_T\}$ for all pairs (i, j).

Thus the computational cost of the filter-based EM algorithm on a sequential machine is higher than that of the smoother-based EM algorithm.

Memory Requirements: In the filter-based EM algorithm, only the filtered variables at each time instant need to be stored to compute the variables at the next time instant. They can then be discarded. The memory required in each iteration is $O(m^4)$ and is independent of the number of observations T.

In comparison, the Kalman smoother-based EM algorithm in [2] and [4] requires $O(m^2T)$ memory per EM iteration since all the Kalman filter covariance matrices R_k, $1 \leq k \leq T$ need to be stored before smoothed covariance matrices can be computed; see [2, (2.12)]. This also involves significant memory read–write overhead costs.

B. Parallel Implementation on Systolic Array Architecture

The following properties of the filter-based EM algorithm make it suitable for vector-processor or systolic-processor implementation.

1) The computation of $a_k^{ij(M)}$, $b_k^{ij(M)}$, and $d_k^{ij(M)}$ for each pair (i, j) is independent of $a_k^{i'j'(M)}$, $b_k^{i'j'(M)}$ and $d_k^{i'j'(M)}$ for any other pair (i', j') for *all* time $k = 0, 1, \cdots$. So all the i, j components of these variables can be computed in parallel on m^2 processors.

Similarly computation of all (i, n) components of $\bar{a}_k^{in}$ and $\bar{b}_k^{in}$ are mutually independent and can be done in parallel.

2) The recursions for $a_k^{(M)}$, $b_k^{(M)}$, $d_k^{(M)}$, and $\bar{a}_k^{in}$ do not explicitly involve the observations. They only involve the Kalman filter variables, μ_{k-1}, σ_k, $R_{k|k-1}$. Notice that μ_k only arises in the term $S_k = \sigma_{k+1}^{-1} R_k^{-1} \mu_k$. This term S_k arises in (33), (34), (36), (37), (39), (40), and (51) and only needs to be computed once for each time k.

Moreover, $d_k^{(M)}$ only involves R_k (and so $R_{k|k-1}$) which itself is independent of the observations and can be computed off-line for a given parameter set θ. Similarly the term $\Sigma_{k+1} = R_{k+1|k}^{-1} A_{k+1} R_k$ arises in (34), (35), (37)–(41), and (52) and can be computed off-line for a given θ.

All the processor blocks used above are required to do a synchronous matrix vector multiplication at every time instant k. Now an $N \times N$ matrix can be multiplied by a N vector in N time units on a N processor systolic array; see [17, pp. 216–220] for details. (Also it can also be done in unit time on N^2 processors).

If τ is the time required for this matrix-vector multiplication, then for a T-point data sequence, the filter-based EM algorithm requires a total of τT time units per EM iteration. In comparison, a parallel implementation of the forward–backward smoother-based EM algorithm requires $2\tau T$ time units per EM iteration because we need a minimum of τT time units to compute the forward variables and another τT units for the backward variables. For large T and a large number of EM iterations, this saving in time is quite considerable.

In addition, unlike the filter-based EM algorithm which has negligible memory requirements, the forward–backward algorithm of the smoother-based EM requires significant memory read–write overhead costs requiring m^2T memory locations to be accessed for the stored forward variables while computing the backward variables.

Finally, the filter-based EM algorithm can be easily implemented in a single instruction multiple data (SIMD) mode on a supercomputer in the vectorization mode or the Connection Machine using FORTRAN 8X. Typically with $m = 10$, we need a total $O(100)$ matrix vector multiplications per time instant. That is we need a total of 10 000 processor units on a Connection Machine, which typically has $2^{16} = 65\,536$ processors.

X. Conclusions and Future Work

We have presented a new class of finite-dimensional filters for linear Gauss–Markov models that includes the Kalman filter as a special case. These filters were then used to derive a filter-based expectation maximization algorithm for computing MLE's of the parameters.

It is possible to derive the filters in continuous-time using similar techniques. This is the subject of a companion paper [18].

It is of interest to apply the results in this paper to recursive parameter estimation. A recursive version of the smoother-based EM algorithm which approximates the smoothed estimates at each time instant by filtered estimates has been proposed by [3] and used for parameter estimation of errors-in-variables models in [4] and [5]. It would be interesting to derive a recursive EM algorithm based on the filter-based EM algorithm developed in this paper. Also the convergence of such a stochastic approximation algorithm and its application in adaptive control could be studied.

Appendix A
Proof of Lemma 3.2

Proof: Suppose $f : R^d \to R$ and $g : R^m \to R$ are arbitrary measurable "test" functions. Then with $\mathbf{E}$ (respectively, $\bar{\mathbf{E}}$) denoting expectation under P (respectively, $\bar{P}$)

$$\mathbf{E}\{g(w_k)f(v_k) \mid \mathcal{G}_{k-1}\} = \frac{\bar{\mathbf{E}}\{\Lambda_k g(w_k)f(v_k) \mid \mathcal{G}_{k-1}\}}{\bar{\mathbf{E}}\{\Lambda_k \mid \mathcal{G}_{k-1}\}} \qquad (81)$$

using a version of Bayes' theorem [11].

$$\begin{aligned}\mathbf{E}\{g(w_k)f(v_k) \mid \mathcal{G}_{k-1}\} =&\bar{\mathbf{E}}\{\lambda_k g(w_k)f(v_k) \mid \mathcal{G}_{k-1}\} = \bar{\mathbf{E}}\left\{\frac{\psi\big(B_k^{-1}(x_k - A_k x_{k-1})\big)}{|B_k|\psi(x_k)}\frac{\phi\big(D_k^{-1}(y_k - C_k x_k)\big)}{|D_k|\phi(y_k)}\right.\\ &\left.\times g\big(B_k^{-1}(x_k - A_k x_{k-1})\big) f\big(D_k^{-1}(y_k - C_k x_k)\big) \,\middle|\, \mathcal{G}_{k-1}\right\}\\ =&\bar{\mathbf{E}}\left\{\frac{\psi\big(B_k^{-1}(x_k - A_k x_{k-1})\big)}{|B_k|\psi(x_k)} g\big(B_k^{-1}(x_k - A_k x_{k-1})\right.\\ &\left.\times\,\bar{\mathbf{E}}\left\{\frac{\phi(D_k^{-1}(y_k - C_k x_k))}{|D_k|\phi(y_k)} f(D_k^{-1}(y_k - C_k x_k)) \,\middle|\, \mathcal{G}_{k-1}, x_k\right\} \,\middle|\, \mathcal{G}_{k-1}\right\}\end{aligned} \tag{83}$$

Now Λ_{k-1} is $\mathcal{G}_{k-1}$ measurable, therefore

$$\mathbf{E}\{g(w_k)f(v_k) \mid \mathcal{G}_{k-1}\} = \frac{\bar{\mathbf{E}}\{\lambda_k g(w_k)f(v_k) \mid \mathcal{G}_{k-1}\}}{\bar{\mathbf{E}}\{\lambda_k \mid \mathcal{G}_{k-1}\}}.$$

However

$$\begin{aligned}&\bar{\mathbf{E}}\{\lambda_k \mid \mathcal{G}_{k-1}\}\\ &\quad= \bar{\mathbf{E}}\left\{\frac{\psi\big(B_k^{-1}(x_k - A_k x_{k-1})\big)}{|B_k|\psi(x_k)}\right.\\ &\qquad\left.\times\frac{\phi\big(D_k^{-1}(y_k - C_k x_k)\big)}{|D_k|\phi(y_k)} \,\middle|\, \mathcal{G}_{k-1}\right\}\\ &\quad= \bar{\mathbf{E}}\left\{\frac{\psi\big(B_k^{-1}(x_k - A_k x_{k-1})\big)}{|B_k|\psi(x_k)}\right.\\ &\qquad\left.\times\,\bar{\mathbf{E}}\left\{\frac{\phi\big(D_k^{-1}(y_k - C_k x_k)\big)}{|D_k|\phi(y_k)} \,\middle|\, \mathcal{G}_{k-1}, x_k\right\} \,\middle|\, \mathcal{G}_{k-1}\right\}.\end{aligned} \tag{82}$$

Notice that the inner conditional expectation is

$$\frac{1}{|D_k|}\int_{R^m} \phi\big(D_k^{-1}(y_k - C_k x_k)\big)\,dy_k = 1.$$

Hence

$$\bar{\mathbf{E}}\{\lambda_k \mid \mathcal{G}_{k-1}\} = \frac{1}{|B_k|}\int_{R^m} \psi\big(B_k^{-1}(x_k - A_k x_{k-1})\big)\,dx_k = 1.$$

Consequently, we have (83), as shown at the top of the page. The inner conditional expectation in (83) is

$$\frac{1}{|D_k|}\int_{R^d} \phi\big(D_k^{-1}(y_k - C_k x_k)\big) f\big(D_k^{-1}(y_k - C_k x_k)\big)\,dy_k.$$

Denoting $v_k = D_k^{-1}(y_k - C_k x_k)$ the above expression is

$$\int_{R^d} \phi(v_k) f(v_k)\,dv_k$$

which is independent of all $x_0, x_1, \cdots, x_{k-1}, y_0, \cdots, y_{k-1}$, that is, it is independent of $\mathcal{G}_{k-1}$. Therefore

$$\begin{aligned}&\mathbf{E}\{g(w_k)f(v_k) \mid \mathcal{G}_{k-1}\}\\ &\quad= \int_{R^d} \phi(v)f(v)\,dv \int_{R^m} \psi(w)g(w)\,dw\end{aligned}$$

and the lemma is proved. □

Appendix B
Derivation of $Q(\theta, \tilde{\theta})$ in (5)

Consider the time-invariant state-space model given by (3), (4) with $\theta = (A, B, C, D)$ denoting a possible set of parameters.

It has been shown in Section III how starting from a measure $\bar{P}$ under which the x_l and v_l are independent and normal, one can construct the measure $P = P(\theta)$, such that under $P = P(\theta)$, the x and y sequences satisfy the dynamics (1) and (2). In fact

$$\left.\frac{\partial P(\theta)}{\partial \bar{P}}\right|_{\mathcal{G}_k} = \Lambda_k(\theta).$$

Suppose $\tilde{\theta} = (\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ is a second set of parameters. Then

$$\left.\frac{\partial P(\tilde{\theta})}{\partial \bar{P}}\right|_{\mathcal{G}_k} = \Lambda_k(\tilde{\theta}).$$

To change from, say, one set of parameters $\tilde{\theta}$ to θ, we must introduce the densities

$$\Gamma_k(\tilde{\theta}, \theta) = \frac{\Lambda_k(\theta)}{\Lambda_k(\tilde{\theta})} = \prod_{l=0}^{k} \gamma_l$$

where

$$\begin{aligned}\gamma_0 &= \frac{|\tilde{D}|}{|D|}\frac{\phi(D^{-1}(y_0 - Cx_0))}{\phi(\tilde{D}^{-1}(y_0 - \tilde{C}x_0))}\\ \gamma_l &= \frac{|\tilde{B}|}{|B|}\frac{\psi(B^{-1}(x_l - Ax_{l-1}))}{\psi(\tilde{B}^{-1}(x_l - \tilde{A}x_{l-1}))}\frac{|\tilde{D}|}{|D|}\frac{\phi(D^{-1}(y_l - Cx_l))}{\phi(\tilde{D}^{-1}(y_l - \tilde{C}x_l))}.\end{aligned}$$

The parameters of our model will be changed from $\tilde{\theta}$ to θ if we set

$$\left.\frac{dP(\theta)}{dP(\tilde{\theta})}\right|_{\mathcal{G}_k} = \Gamma_k(\tilde{\theta}, \theta).$$

In this case

$$\begin{aligned}\log\left.\frac{dP(\theta)}{dP(\tilde{\theta})}\right|_{\mathcal{G}_k} &= -k\log|B| - (k+1)\log|D|\\ &\quad - \frac{1}{2}\sum_{l=1}^{k}(x_l - Ax_{l-1})'B^{-2}(x_l - Ax_{l-1})\\ &\quad - \frac{1}{2}\sum_{l=0}^{k}(y_l - Cx_l)'D^{-2}(y_l - Cx_l) + R(\tilde{\theta})\end{aligned}$$

where $R(\tilde{\theta})$ does not involve any of the parameters θ.

Then evaluating $Q(\theta, \tilde{\theta}) = \mathbf{E}_{\tilde{\theta}}\{\log \frac{dP(\theta)}{dP(\tilde{\theta})} \mid \mathcal{Y}_T\}$ for a fixed positive integer T yields (5).

References

[1] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *J. Time Series Anal.*, vol. 3, no. 4, pp. 253–264, 1982.

[2] D. Ghosh, "Maximum likelihood estimation of the dynamic shock-error model," *J. Econometrics*, vol. 41, no. 1, pp. 121–143, May 1989.

[3] D. M. Titterington, "Recursive parameter estimation using incomplete data," *J. R. Statistical Society B.*, vol. 46, pp. 257–267, 1984.

[4] E. Weinstein, A. V. Oppenheim, M. Feder, and J. R. Buck, "Iterative and sequential algorithms for multisensor signal enhancement," *IEEE Signal Processing*, vol. 42, pp. 846–859, Apr. 1994.

[5] V. Krishnamurthy, "On-line estimation of dynamic shock-error models," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 1129–1134, 1994.

[6] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

[7] I. Ziskind and D. Hertz, "Maximum likelihood localization of narrow-band autoregressive sources via the EM algorithm," *IEEE Trans. Signal Processing*, vol. 41, no. 8, pp. 2719–2723, Aug. 1993.

[8] L. Breiman, "Probability," *Classics in Applied Mathematics*, vol. 7. Philadelphia, PA: SIAM, 1992.

[9] R. J. Elliott, "Exact adaptive filters for Markov chains observed in Gaussian noise," *Automatica*, vol. 30, no. 9, pp. 1399–1408, Sept. 1994.

[10] P. E. Caines, *Linear Stochastic Systems*. New York: Wiley, 1988.

[11] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov Models: Estimation and Control*. New York: Springer-Verlag, 1995.

[12] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. New York: McGraw Hill, 1984.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society, B*, vol. 39, pp. 1–38, 1977.

[14] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Annals of Statistics*, vol. 11, pp. 95–103, 1983.

[15] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. New York: Academic, 1970.

[16] P. Billingsley, *Probability and Measure*, 2nd ed. New York: Wiley, 1986.

[17] E. V. Krishnamurthy, *Parallel Processing: Principles and Practice*. New York: Addison Wesley, 1989.

[18] R. J. Elliott and V. Krishnamurthy, "New finite dimensional filters for estimation of continuous-time linear Gaussian systems," *SIAM J. Contr. Optim.*, vol. 35, no. 6, pp. 1908–1923, Nov. 1997.

Robert J. Elliott received the Bachelors and Masters degrees from Oxford University and the Ph.D. and D.Sc. from Cambridge University.

He is currently a Professor in the Department of Mathematical Sciences, University of Alberta. He has held positions at Newcastle, Yale, Oxford, Warwick, Hull, and Alberta, and visiting positions in Toronto, Northwestern, Kentucky, Brown, Paris, Denmark, Hong Kong, and Australia. He has authored over 250 papers and five books, in particular, *Hidden Markov Models: Estimation and Control* (New York: Springer-Verlag, 1995) with L. Aggoun and J. Moore. His work in recent years has investigated stochastic processes in engineering and finance.

Vikram Krishnamurthy (S'90–M'91) was born in India in 1966. He received the bachelor's degree in electrical engineering from the University of Auckland, New Zealand, in 1988, and doctoral degree from the Australian National University, Canberra in 1992.

He is currently an Associate Professor in the Department of Electrical Engineering, University of Melbourne. His research interests include time-series analysis, stochastic filtering theory, and statistical signal processing in communication systems.