Systems & Control Letters 119 (2018) 92–100

Contents lists available at ScienceDirect

Systems & Control Letters

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

Suboptimal linear estimation for continuous–discrete bilinear systems

Xue Luo [a], Xiuqiong Chen [b,1], Stephen S.-T. Yau [b,*]

[a] School of Mathematics and Systems Science, Beihang University, Beijing 100191, China
[b] Department of Mathematical Sciences, Tsinghua University, Beijing, 100084, China

ARTICLE INFO

Article history:
Received 26 September 2017
Received in revised form 19 June 2018
Accepted 16 July 2018

Dedicated to Professor T. Duncan on the occasion of his 75th birthday.

Keywords:
Nonlinear filtering
Bilinear systems
Carleman approach
The extended Kalman filter

ABSTRACT

In this paper we derive a suboptimal estimation for continuous–discrete bilinear systems. One of the motivations of this work is that the bilinear system has the simplest structure in the nonlinear class in some sense. Similar to the Kalman filter, our algorithm includes prediction and updating step. We show rigorously that our algorithm gives an unbiased estimate, the a-priori estimate approaches to the conditional expectation exponentially fast, and the posterior estimate minimizes the conditional variance error in the linear space spanned by the a-priori estimate and the innovation. Our algorithm is also applicable to solve the nonlinear filtering problems. The efficiency of our method is illustrated by the cubic sensor problem and Lorenz system with discrete observation. The results have been compared with the extended Kalman filter and the unscented Kalman filter.

1. Introduction

How to get the instantaneous and accurate estimation of the states of a stochastic system from the polluted measurements by the noise is of central importance in engineering and this is also the central problem in the field of filtering. A continuous–discrete filtering problem is modeled by the following Itô stochastic differential equation:

$$\begin{cases} dx(t) = f(x(t))dt + g(x(t))dv(t) \\ y(t_k) = h(x(t_k)) + w(t_k), \end{cases} \tag{1.1}$$

where $v(t)$ is Brownian motion with proper dimension, $x(t) \in \mathbb{R}^n$ is the state, $0 = t_0 < t_1 < \cdots < t_K = T$, $y(t_k) \in \mathbb{R}^m$ is the measurement, t_k, $k = 1, 2, \ldots, K$ are instants when the measurements arrive and $w(t_k) \in \mathbb{R}^m$ is white noise. When the function $f(x)$ and $h(x)$ are linear functions of x and $g(x)$ is constant, we call (1.1) a linear filtering problem and its study can be traced back to early 1960s when Kalman [1], Kalman and Bucy [2] published two most influential papers and proposed the classical Kalman filter and Kalman–Bucy filter. We refer the readers to the book [3] for excellent introduction to filtering theory. Though the linear filtering problem is completely solved in [1,2,4], the nonlinear filtering (NLF) problems are much more complicated and important in applications since most practical models are nonlinear.

One class of methods to solve NLF problems is the so-called global approaches which try to find out the conditional density function of the states by solving the Duncan–Mortensen–Zakai (DMZ) equation [5–7]. Based on the DMZ equation, more research articles follow this direction such as [8–13]. Numerical methods to solve this problem can also be found such as in [14].

Another class of methods to solve the NLF problems is referred as local approaches, which construct suboptimal filters in some sense. There are many approximate methods including unscented Kalman filter (UKF) [15,16], ensemble Kalman filter [17], particle filter [18] and the most widely used extended Kalman filter (EKF) [3,19], which is basically the Kalman–Bucy filter applied to a linearized system. However, EKF can only perform well if the initial estimation error and the disturbing noises are small enough due to its local nature.

Continuous–discrete filter, which is for stochastic differential systems with sampled measurements, is also of great significance and has many applications such as in tracking and finance since the measurements always come in discretely. There has been increasing interest in this system and many continuous–discrete filters can be found in the literatures, such as continuous–discrete EKF [3], continuous–discrete UKF [15], continuous–discrete Gaussian filter [20] and continuous–discrete cubature Kalman filter [21]. And the comparison of these different methods can refer [22].

Our motivations to study the bilinear system are two folds: on the one hand, many important processes, not only in engineering but also in socio-economics, biology and ecology, may be modeled by bilinear systems [23]. On the other hand, the bilinear structure

* Corresponding author.
E-mail addresses: xluo@buaa.edu.cn (X. Luo), cxq14@mails.tsinghua.edu.cn (X. Chen), yau@uic.edu (S.S.-T. Yau).
[1] X. Chen is the co-first author.

https://doi.org/10.1016/j.sysconle.2018.07.005
0167-6911/

seems to be the simplest and closest one to the linear one among all the nonlinearities. Thus, some well-established techniques can be extended to bilinear systems [24]. The estimation theory for bilinear system can also be used to solve NLF problems. For example, the nonlinear analytical system can be approximated by a bilinear system using Carleman approach [25].

Notice that [24] only deals with the continuous–continuous systems. Recently, the first and the last author of this paper [26,27] proposed a novel algorithm for solving the continuous NLF problems based on the idea in [24]. In this paper we derive a suboptimal filter for the continuous–discrete bilinear systems. Compared with the work of Cacace and his collaborators, we consider the filter rather than state predictor [28] and the bilinear system (3.1) in our algorithm is more general than that in [29]. We call the estimate obtained in this paper suboptimal linear estimate (SLE). Similar to EKF, our algorithm consists of two steps including predicting and updating. We call the estimate after prediction the a-priori estimate, while that after updating the posterior estimate. The suboptimality of our algorithm in the following sense: essentially, we show that under some mild conditions SLE has the following properties:

1. Both the a-priori and the posterior estimates are unbiased;
2. The a-priori estimate approaches to the conditional expectation exponentially fast;
3. The posterior estimate minimizes the conditional variance error in a linear space.

This paper is organized as follows. Our algorithm is described in Section 2.1. The suboptimality of SLE has been shown rigorously in Section 2.2. Section 3 presents the application of our algorithm to representative NLF problems, where we compare the performance of the proposed filter with EKF and UKF. We arrive at the conclusion in Section 4.

2. Suboptimal algorithm

The bilinear continuous–discrete system considered in probability space $(\Omega, \mathscr{F}, P)$ is as follows:

$$\begin{cases} dX(t) = \mathbf{A}X(t)dt + \mathbf{N}dt + \sum_{j=1}^{b}(\mathbf{B}_j X(t) + \mathbf{F}_j)dW_j(t),\ t \in [0, T], \\ Y(t_k) = \mathbf{C}X(t_k) + \mathbf{D} + \sum_{j=1}^{b}\mathbf{G}_j V_j(t_k),\ k = 0, 1, \ldots, K, \end{cases} \tag{2.1}$$

where $0 = t_0 < t_1 < \cdots < t_K = T$, $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{N} \in \mathbb{R}^{n\times 1}$, $\mathbf{B}_j \in \mathbb{R}^{n\times n}$, $\mathbf{F}_j \in \mathbb{R}^{n\times 1}$, $\mathbf{C} \in \mathbb{R}^{m\times n}$, $\mathbf{D} \in \mathbb{R}^{m\times 1}$, $\mathbf{G} \in \mathbb{R}^{m\times 1}$ are constant matrices. $X(t) \in \mathbb{R}^n$ is the state with the initial value X_0 whose mean is $\bar{X}_0$ and covariance matrix is $\bar{P}_0$, $Y(t_k) \in \mathbb{R}^m$ is discrete measurement, $V_j(t_k) \sim \mathcal{N}(0, R_j(t_k))$, $R_j(t_k) \in \mathbb{R}$, $k = 0, 1, \ldots, K$, are independent one-dimensional white noises and $W_j(t), j = 1, \ldots, b$, are independent standard Brownian motions. Let $\mathscr{F}_{t_k}$ be the σ-field generated by the observations, i.e. $\mathscr{F}_{t_k} \triangleq \sigma\{Y(t_0), Y(t_1), \ldots, Y(t_k)\}$. Kronecker algebra is used for concise notation and derivation. Its properties can be found in [30].

Recall that the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with finite second moment, with scalar product $\langle x, y\rangle = E[x^T y]$ and norm $\|x\| := E^{1/2}[x^T x]$ is a Hilbert space, denoted as $L^2(\Omega, \mathscr{F}, \mathbb{P})$. Suppose that the state $X(t) \in L^2(\Omega, \mathscr{F}, \mathbb{P})$. In our algorithm, we shall obtain a linear recursive estimate in the similar fashion of EKF. As explained in [28], the prediction of the state on the observation history is indeed a random variable. After the approximation of the conditional expectation of the nonlinear drift term $f(x_t)$ coarsely in EKF, i.e. $E[f(x_t)|\mathscr{F}_{t_{k-1}}] \approx f(E[x_t|\mathscr{F}_{t_{k-1}}])$, for $t > t_{k-1}$, it makes the state estimate satisfy an deterministic ordinary differential equation. This is the essential reason why all the estimates in our algorithm will be treated in a deterministic way. Let us clearly define the linear recursive estimate of $X(t)$ based on the observation history $\{Y(t_0), Y(t_1), \ldots, Y(t_{k-1})\}$ first:

Definition 2.1. We call $\hat{X}(t_k|t_k)$ the linear recursive estimate of $X(t_k)$ based on the observation $\{Y(t_0), Y(t_1), \ldots, Y(t_k)\}$, if

1. The a-priori estimate, denoted as $\hat{X}(t|t_{k-1})$, $t \in [t_{k-1}, t_k]$, is linearly dependent of the previous posterior estimate $\hat{X}(t_{k-1}|t_{k-1})$, i.e.
$$\hat{X}(t|t_{k-1}) = H_1(t)\hat{X}(t_{k-1}|t_{k-1}) + H_2(t), \tag{2.2}$$
where H_1 and H_2 are matrices of proper dimensions;
2. The posterior estimate $\hat{X}(t_k|t_k)$ lives in the linear space spanned by 1, the a-priori linear estimate $\hat{X}(t_k|t_{k-1})$ and the innovation $Y(t_k) - \hat{Y}(t_k|t_{k-1})$, where $\hat{Y}(t_k|t_{k-1}) = C\hat{X}(t_k|t_{k-1}) + D$. That is,
$$\hat{X}(t_k|t_k) = H_3\hat{X}(t_k|t_{k-1}) + H_4(Y(t_k) - \hat{Y}(t_k|t_{k-1})) + H_5, \tag{2.3}$$
where H_3, H_4 and H_5 are constant matrices of proper dimensions.

2.1. Algorithm

Our algorithm consists of two steps: prediction and updating. Throughout the process, we assume that

(As) $\mathbf{A}$ and $\mathbf{A}_{ex}$ are Hurwitz, where $\mathbf{A}_{ex} := \sum_{l=1}^{b}(\mathbf{B}_l \otimes \mathbf{B}_l) + I_n \otimes \mathbf{A} + \mathbf{A} \otimes I_n$.

We state our algorithm first:

(Al-1) Prediction In the interval $[t_{k-1}, t_k)$, the a-priori estimate $\hat{X}(t|t_{k-1})$ of $X(t)$ based on data $\{Y(t_0), Y(t_1), \ldots, Y(t_{k-1})\}$ satisfies

$$\dot{\hat{X}}(t|t_{k-1}) = \mathbf{A}\hat{X}(t|t_{k-1}) + \mathbf{N}, \tag{2.4}$$

$$\begin{aligned} \dot{Q}(t|t_{k-1}) =& \mathbf{A}Q(t|t_{k-1}) + Q(t|t_{k-1})\mathbf{A}^T \\ &+ \sum_{j=1}^{b}\Big[\mathbf{B}_j Q(t|t_{k-1})\mathbf{B}_j^T \\ &+ \left(\mathbf{B}_j\hat{X}(t|t_{k-1}) + \mathbf{F}_j\right) \\ &\times \left(\mathbf{B}_j\hat{X}(t|t_{k-1}) + \mathbf{F}_j\right)^T\Big], \end{aligned} \tag{2.5}$$

with the initial value $\hat{X}(t_{k-1}|t_{k-1})$ and $Q(t_{k-1}|t_{k-1})$ from previous updating, $\hat{X}(t_0|t_0) := \bar{X}_0$, and $Q(t_0|t_0) := \bar{P}_{t_0}$.

(Al-2) Updating The posterior estimate $\hat{X}(t_k|t_k)$ of $X(t_k)$ based on the observation history $\mathscr{F}_{t_k}$ satisfies

$$\hat{X}(t_k|t_k) = \hat{X}(t_k|t_{k-1}) + K_k\Big[Y(t_k) - \hat{Y}(t_k|t_{k-1})\Big], \tag{2.6}$$

with $\hat{Y}(t_k|t_{k-1}) = \mathbf{C}\hat{X}(t_k|t_{k-1}) + \mathbf{D}$, and the gain function K_k is given by

$$\begin{aligned} K_k =& Q(t_k|t_{k-1})\mathbf{C}^T \\ &\times \left[\mathbf{C}Q(t_k|t_{k-1})\mathbf{C}^T + \sum_{j=1}^{b}\mathbf{G}_j R_j(t_k)(\mathbf{G}_j)^T\right]^{-1}. \end{aligned} \tag{2.7}$$

Meanwhile, the matrix $Q(t_k|t_k)$ is updated by

$$Q(t_k|t_k) = (I_n - K_k\mathbf{C})\,Q(t_k|t_{k-1}). \tag{2.8}$$

where I_n is the identity matrix of dimension $n \times n$.

Remark 2.1. The matrix $Q(t|t_{k-1})$ plays the role of the conditional variance

$$P(t|t_{k-1}) := E\left[(X(t)-\hat{X}(t|t_{k-1}))(X(t)-\hat{X}(t|t_{k-1}))^T\Big|\mathscr{F}_{t_{k-1}}\right]$$

in the algorithm. However, in general, $Q(t|t_{k-1}) \neq P(t|t_{k-1})$, $t \in (t_{k-1}, t_k)$, even if with the same initial value at $t = t_{k-1}$. This can be seen from the delicate analysis of e_{QP} in the proof of Proposition 2.1.

2.2. Suboptimality

Under Assumption (As), our algorithm (Al-1)–(Al-2) gives a suboptimal linear estimate (SLE) in the following sense:

(S-1) If $\hat{X}(t_0|t_0)$ is unbiased, so is $\hat{X}(t_k|t_k)$ in the usual sense. That is,

$$E\left(\hat{X}(t_k|t_k)-X(t_k)\right)=0, \tag{2.9}$$

for $k = 1, \ldots, K$.

(S-2) The a-priori estimate $\hat{X}(t|t_{k-1})$, for $t \in [t_{k-1}, t_k]$ approaches the conditional expectation $E\left(X(t)|\mathscr{F}_{t_{k-1}}\right)$ component-wisely and exponentially fast with respect to t. Also $Q(t|t_{k-1})$ approaches the conditional variance $P(t|t_{k-1})$ component-wisely and exponentially fast with respect to t.

(S-3) The posterior estimate $\hat{X}(t_k|t_k)$ minimizes the conditional variance error $\operatorname{tr} P(t_k|t_k)$ in the linear space spanned by $\{1, \hat{X}(t_k|t_{k-1}), Y(t_k) - \hat{Y}(t_k|t_{k-1})\}$, where $\operatorname{tr}\star$ denotes the trace of $\star$.

2.2.1. Proof of (S-1)

Before we proceed, we need the following lemma.

Lemma 2.1 (*Proposition 2.10, [31]*)**.** *If $f(t) \in L^2$ and $f(t)$ is adapted with respect to (w.r.t.) $\tilde{\mathscr{F}}_t \triangleq \sigma\{B_\tau, \tau \le t\}$, where $\{B_\tau\}$ is a Brownian motion, then $I(t) = \int_{t_0}^t f(t)dB_t$ is a martingale w.r.t. $\tilde{\mathscr{F}}_t$ and we have*

$$E\left[\int_{t_0}^t f(t)dB_t\Big|\tilde{\mathscr{F}}_{t_0}\right]=0. \tag{2.10}$$

If $f(t) \in L^2$,and $f(t)$ is adapted w.r.t. $\tilde{\mathscr{F}}_t$, then we have

$$E\left[\int_s^t f(u)du\Big|\tilde{\mathscr{F}}_s\right]=\int_s^t E[f(u)|\tilde{\mathscr{F}}_s]du. \tag{2.11}$$

Proof of (S-1) (*By Induction*)**.** It is sufficient to show that if $\hat{X}(t_{k-1}|t_{k-1})$ is unbiased, so is $\hat{X}(t_k|t_k)$, for $k = 1, 2, \ldots, K-1$.

Solving the first equation in (2.1), we can obtain [32]

$$\begin{aligned}X(t) =& e^{\mathbf{A}(t-t_{k-1})}X(t_{k-1})+\left(e^{A(t-t_{k-1})}-I\right)\mathbf{A}^{-1}\mathbf{N}\\ &+\sum_{j=1}^b\int_{t_{k-1}}^t e^{\mathbf{A}(t-\tau)}(\mathbf{B}_jX(\tau)+\mathbf{F}_j)dW_j(\tau),\end{aligned} \tag{2.12}$$

since $\mathbf{A}^{-1}$ exists by Assumption (As). We claim that the a-priori estimate $\hat{X}(t_k|t_{k-1})$ is unbiased. In (Al-1), the solution of (2.4) is in the form (2.2) with

$$H_1(t)=e^{\mathbf{A}(t-t_{k-1})},\quad H_2(t)=\left(e^{\mathbf{A}(t-t_{k-1})}-I\right)\mathbf{A}^{-1}\mathbf{N}. \tag{2.13}$$

It is easy to see that

$$\begin{aligned}E(X(t_k)) \overset{(2.12),(2.13)}{=}& H_1(t_k)E(X(t_{k-1}))+H_2(t_k)\\ =& H_1(t_k)E(\hat{X}(t_{k-1}|t_{k-1}))+H_2(t_k)=E(\hat{X}(t_k|t_{k-1})),\end{aligned} \tag{2.14}$$

where the second equality follows from the assumption that $\hat{X}(t_{k-1}|t_{k-1})$ is unbiased, for all $k = 1, 2, \ldots, K$. Next, we show that the posterior estimate $\hat{X}(t_k|t_k)$ is unbiased. In (Al-2), we have

$$\begin{aligned}E\left(\hat{X}(t_k|t_k)\right) \overset{(2.6)}{=}& E\left(\hat{X}(t_k|t_{k-1})\right)+K_kE\left(Y(t_k)-\hat{Y}(t_k|t_{k-1})\right)\\ =& E(X(t_k))+K_k\mathbf{C}E(X(t_k)-\hat{X}(t_k|t_{k-1})) \overset{(2.14)}{=} E(X(t_k)).\end{aligned}$$

That is, $\hat{X}(t_k|t_k)$ is unbiased. This completes our induction. □

2.2.2. Proof of (S-2)

Let us first derive the evolution equation of the conditional expectation

$$\tilde{X}(t|t_{k-1}) := E[X(t)|\mathscr{F}_{t_{k-1}}],$$

and the conditional covariance matrix

$$\tilde{P}(t|t_{k-1}) := E\left[\left(X(t)-\tilde{X}(t|t_{k-1})\right)\left(X(t)-\tilde{X}(t|t_{k-1})\right)^T\Big|\mathscr{F}_{t_{k-1}}\right].$$

Lemma 2.2. *In the time interval $[t_{k-1}, t_k)$, suppose that the state propagates according to (2.1). Then the conditional expectation and covariance matrix evolves according to the following equations:*

$$\dot{\tilde{X}}(t|t_{k-1}) = \mathbf{A}\tilde{X}(t|t_{k-1})+\mathbf{N}, \tag{2.15}$$

$$\begin{aligned}\dot{\tilde{P}}(t|t_{k-1}) =& \mathbf{A}\tilde{P}(t|t_{k-1})+\tilde{P}(t|t_{k-1})\mathbf{A}^T\\ &+\sum_{j=1}^b E\left[(\mathbf{B}_jX(t)+\mathbf{F}_j)(\mathbf{B}_jX(t)+\mathbf{F}_j)^T\Big|\mathscr{F}_{t_{k-1}}\right]\\ =& \mathbf{A}\tilde{P}(t|t_{k-1})+\tilde{P}(t|t_{k-1})\mathbf{A}^T\\ &+\sum_{j=1}^b\Big[\mathbf{B}_j\tilde{P}(t|t_{k-1})\mathbf{B}_j^T+(\mathbf{B}_j\tilde{X}(t|t_{k-1})+\mathbf{F}_j)\\ &\times(\mathbf{B}_j\tilde{X}(t|t_{k-1})+\mathbf{F}_j)^T\Big].\end{aligned} \tag{2.16}$$

We omit the proof of this lemma, since it is standard. The following proposition shows that the difference between the a-priori estimate in our algorithm and the conditional expectation vanishes fast.

Proposition 2.1. *In the interval $t \in [t_{k-1}, t_k]$, let us denote $e_X(t|t_{k-1}) := \hat{X}(t|t_{k-1}) - \tilde{X}(t|t_{k-1})$ and $e_{QP} := Q(t|t_{k-1}) - P(t|t_{k-1})$. Then under Assumption (As), we have*

$$e_X(t|t_{k-1})\to 0, \tag{2.17}$$

$$e_{QP}(t|t_{k-1})\to 0, \tag{2.18}$$

component-wisely and exponentially fast, as $t \to \infty$.

Proof. The proof of (2.17) is straightforward. Due to the linearity of (2.4) and (2.15), we have

$$\dot{e}_X(t|t_{k-1}) = \mathbf{A}e_X(t|t_{k-1}), \tag{2.19}$$

with the initial condition $e_X(t_{k-1}|t_{k-1})$. It can be solved explicitly that

$$e_X(t|t_{k-1}) = e^{\mathbf{A}(t-t_{k-1})}e_X(t_{k-1}|t_{k-1}). \tag{2.20}$$

Thus, (2.17) follows immediately from Lemma A.3, since $\mathbf{A}$ is Hurwitz. Similar to the derivation of (2.16), one obtains that

$$\begin{aligned}\dot{P}(t|t_{k-1}) =& \mathbf{A}P(t|t_{k-1})+P(t|t_{k-1})\mathbf{A}^T\\ &+\sum_{j=1}^b E\left[(\mathbf{B}_jX(t)+\mathbf{F}_j)(\mathbf{B}_jX(t)+\mathbf{F}_j)^T\Big|\mathscr{F}_{t_{k-1}}\right].\end{aligned} \tag{2.21}$$

Comparing (2.21) and (2.15), it is easy to see that $e_{P\tilde{P}}(t|t_{k-1}) := P(t|t_{k-1}) - \tilde{P}(t|t_{k-1})$ satisfies

$$\dot{e}_{P\tilde{P}}(t|t_{k-1}) = \mathbf{A}e_{P\tilde{P}}(t|t_{k-1})+e_{P\tilde{P}}(t|t_{k-1})\mathbf{A}^T. \tag{2.22}$$

Vectorizing (2.22) yields that

$$\frac{d}{dt}\text{vec}\left(e_{P\tilde{P}}(t|t_{k-1})\right) = \text{vec}\left(\dot{e}_{P\tilde{P}}(t|t_{k-1})\right) = (I_n\otimes\mathbf{A}+\mathbf{A}\otimes I_n)\text{vec}\left(e_{P\tilde{P}}(t|t_{k-1})\right),$$

where $\text{vec}(\circ_{m\times n})$ is the $mn\times 1$ column vector obtained by stacking the columns of the matrix $\circ$ on top of one another. Thus, we have

$$\begin{aligned}\text{vec}\left(e_{P\tilde{P}}(t|t_{k-1})\right) =& e^{(I_n\otimes\mathbf{A}+\mathbf{A}\otimes I_n)(t-t_{k-1})}\text{vec}\left(e_{P\tilde{P}}(t_{k-1}|t_{k-1})\right) \\ =& e^{(I_n\otimes\mathbf{A})(t-t_{k-1})}e^{(\mathbf{A}\otimes I_n)(t-t_{k-1})}\text{vec}\left(e_{P\tilde{P}}(t_{k-1}|t_{k-1})\right) \\ \to& 0_{n^2\times 1},\end{aligned} \tag{2.23}$$

component-wisely and exponentially by Lemma A.3. The second equality follows from the fact that $e^{A+B}=e^Ae^B$, if A and B commute with each other.

Next, we look at the difference between $Q(t|t_{k-1})$ and $\tilde{P}(t|t_{k-1})$. Let us denote $e_{Q\tilde{P}}(t|t_{k-1}) := Q(t|t_{k-1})-\tilde{P}(t|t_{k-1})$. Subtracting (2.16) from (2.5), we get

$$\begin{aligned}\dot{e}_{Q\tilde{P}}(t|t_{k-1}) =& \mathbf{A}e_{Q\tilde{P}}(t|t_{k-1})+e_{Q\tilde{P}}(t|t_{k-1})\mathbf{A}^T+\sum_{j=1}^b\mathbf{B}_je_{Q\tilde{P}}(t|t_{k-1})\mathbf{B}_j^T \\ &+\sum_{j=1}^b\Big[(\mathbf{B}_j\hat{X}(t|t_{k-1})+\mathbf{F}_j)(\mathbf{B}_j\hat{X}(t|t_{k-1})+\mathbf{F}_j)^T \\ &-(\mathbf{B}_j\tilde{X}(t|t_{k-1})+\mathbf{F}_j)(\mathbf{B}_j\tilde{X}(t|t_{k-1})+\mathbf{F}_j)^T\Big].\end{aligned} \tag{2.24}$$

Let us denote the last summation on the right-hand side of (2.24) as I_1. According to (2.4), $\hat{X}(t|t_{k-1})$ can be solved explicitly, i.e.

$$\begin{aligned}\hat{X}(t|t_{k-1}) =& e^{\mathbf{A}(t-t_{k-1})}\hat{X}(t_{k-1}|t_{k-1})-\mathbf{A}^{-1}\left(I-e^{\mathbf{A}(t-t_{k-1})}\right)\mathbf{N} \\ =& e^{\mathbf{A}(t-t_{k-1})}\left(\hat{X}(t_{k-1}|t_{k-1})+\mathbf{A}^{-1}\mathbf{N}\right)-\mathbf{A}^{-1}\mathbf{N},\end{aligned} \tag{2.25}$$

so does $\tilde{X}(t|t_{k-1})$, with the same expression above replacing $\hat{X}$ by $\tilde{X}$. Substituting (2.25) to I_1 and suppressing the notation $(t_{k-1}|t_{k-1})$ in $\hat{X}(t_{k-1}|t_{k-1})$ and $\tilde{X}(t_{k-1}|t_{k-1})$ in the sequel, it yields that

$$\begin{aligned}I_1 =& \sum_{j=1}^b\left[\left(\mathbf{B}_je^{\mathbf{A}(t-t_{k-1})}\bar{\hat{X}}+\bar{\mathbf{F}}_j\right)\left(\mathbf{B}_je^{\mathbf{A}(t-t_{k-1})}\bar{\hat{X}}+\bar{\mathbf{F}}_j\right)^T\right. \\ &\left.-\left(\mathbf{B}_je^{\mathbf{A}(t-t_{k-1})}\bar{\tilde{X}}+\bar{\mathbf{F}}_j\right)\left(\mathbf{B}_je^{\mathbf{A}(t-t_{k-1})}\bar{\tilde{X}}+\bar{\mathbf{F}}_j\right)^T\right] \\ =& \sum_{j=1}^b\mathbf{B}_je^{\mathbf{A}(t-t_{k-1})}\left(\bar{\hat{X}}\bar{\hat{X}}^T-\bar{\tilde{X}}\bar{\tilde{X}}^T\right)e^{\mathbf{A}^T(t-t_{k-1})}\mathbf{B}_j^T \\ &+\sum_{j=1}^b\mathbf{B}_je^{\mathbf{A}(t-t_{k-1})}\left(\bar{\hat{X}}-\bar{\tilde{X}}\right)\bar{\mathbf{F}}_j^T+\sum_{j=1}^b\bar{\mathbf{F}}_j\left(\bar{\hat{X}}-\bar{\tilde{X}}\right)^Te^{\mathbf{A}^T(t-t_{k-1})}\mathbf{B}_j^T \\ =:& II_1+II_2+II_3,\end{aligned}$$

where $\bar{\hat{X}}=\hat{X}+\mathbf{A}^{-1}\mathbf{N}$, $\bar{\tilde{X}}=\tilde{X}+\mathbf{A}^{-1}\mathbf{N}$ and $\bar{\mathbf{F}}_j=-\mathbf{B}_j\mathbf{A}^{-1}\mathbf{N}+\mathbf{F}_j$. Vectorizing (2.24), one obtains that

$$\frac{d}{dt}\text{vec}\left(e_{Q\tilde{P}}(t|t_{k-1})\right)=\mathbf{A}_{ex}\text{vec}\left(e_{Q\tilde{P}}(t|t_{k-1})\right)+\mathbf{N}_{ex}(t), \tag{2.26}$$

where $\mathbf{A}_{ex}$ is given in Assumption (As) and $\mathbf{N}_{ex}(t)=\text{vec}(I_1(t))$. The solution of (2.26) can be expressed in the integral form:

$$\begin{aligned}\text{vec}\left(e_{Q\tilde{P}}(t|t_{k-1})\right) =& e^{\mathbf{A}_{ex}(t-t_{k-1})}\text{vec}\left(e_{Q\tilde{P}}(t_{k-1}|t_{k-1})\right) \\ &+e^{\mathbf{A}_{ex}t}\int_{t_{k-1}}^t e^{-\mathbf{A}_{ex}\tau}\text{vec}(II_1+II_2+II_3)d\tau.\end{aligned} \tag{2.27}$$

The first term on the right-hand side of (2.27) tends to 0 component-wisely and exponentially by Lemma A.3 and $\mathbf{A}_{ex}$ is Hurwitz. In the sequel, we shall analyze the integral on the right-hand side of (2.27) one-by-one. Let us first consider the term containing $\text{vec}(II_2)$:

$$\begin{aligned}&e^{\mathbf{A}_{ex}t}\int_{t_{k-1}}^t e^{-\mathbf{A}_{ex}\tau}\text{vec}(II_2)d\tau \\ =&\sum_{j=1}^b e^{\mathbf{A}_{ex}t}\int_{t_{k-1}}^t e^{-\mathbf{A}_{ex}\tau}\text{vec}\left(\mathbf{B}_je^{\mathbf{A}(\tau-t_{k-1})}e_X\bar{\mathbf{F}}_j^T\right)d\tau \\ =&\sum_{j=1}^b\int_{t_{k-1}}^t e^{\mathbf{A}_{ex}(t-\tau)}\left[(e^{\mathbf{A}(\tau-t_{k-1})}e_X\bar{\mathbf{F}}_j^T)^T\otimes I_n\right]\text{vec}\left(\mathbf{B}_j\right)d\tau \\ =&\sum_{j=1}^b\left\{\int_{t_{k-1}}^t e^{\mathbf{A}_{ex}(t-\tau)}\left[\bar{\mathbf{F}}_je_X^T\otimes I_n\right]e^{\left(\mathbf{A}^T\otimes I_n\right)(\tau-t_{k-1})}d\tau\right\}\text{vec}\left(\mathbf{B}_j\right),\end{aligned} \tag{2.28}$$

where $e_X=\bar{\hat{X}}-\bar{\tilde{X}}=\hat{X}-\tilde{X}$. The last equality follows from the fact that $e^{\mathbf{A}}\otimes I_n=\sum_{k=0}^\infty\frac{\mathbf{A}^k}{k!}\otimes I_n=\sum_{k=0}^\infty\frac{(\mathbf{A}\otimes I_n)^k}{k!}=e^{\mathbf{A}\otimes I_n}$. For short, let us denote $F_j:=\bar{\mathbf{F}}_je_X^T\otimes I_n$ in the sequel. Vectorizing the integral in (2.28), we have

$$\begin{aligned}&\text{vec}\left[\int_{t_{k-1}}^t e^{\mathbf{A}_{ex}(t-\tau)}F_je^{\left(\mathbf{A}^T\otimes I_n\right)(\tau-t_{k-1})}d\tau\right] \\ =&\left[\int_{t_{k-1}}^t e^{(\mathbf{A}\otimes I_n)(\tau-t_{k-1})}\otimes e^{\mathbf{A}_{ex}(t-\tau)}d\tau\right]\text{vec}(F_j) \\ \overset{(A.3)}{=}&\text{vec}^{-1}_{n^4\times n^4}\left\{\left(P|_{J_M^*}\right)_R^{-1}J_M^{*\,-1}\left\{P\,\text{vec}\left[I_{n^2}\otimes\left(e^{\mathbf{A}_{ex}(t-t_{k-1})}\mathbf{A}_{ex}^{-1}\right)\right.\right.\right. \\ &\left.\left.\left.-e^{(\mathbf{A}\otimes I_n)(t-t_{k-1})}\otimes\mathbf{A}_{ex}^{-1}\right]\right\}\Big|_{J_M^*}\right\} \\ &\cdot\text{vec}(F_j),\end{aligned} \tag{2.29}$$

where $\text{vec}^{-1}_{m\times n}$ is the inverse operator of vec such that $\text{vec}^{-1}_{m\times n}(\text{vec}(\circ_{m\times n}))=\circ_{m\times n}$. The notations J_M^* and $\circ|_{J_M^*}$ are specified in Lemma A.2, with $M=I_{n^8}-\mathbf{A}^T\otimes I_{n^5}\otimes\mathbf{A}_{ex}^{-1}$ in this term. The conditions in Lemma A.2 are satisfied, since $\log\left(e^{\mathbf{A}\otimes I_n}\right)=\mathbf{A}\otimes I_n$ and $\mathbf{A}_{ex}$ is invertible by Assumption (As). According to Lemma A.3, $e^{\mathbf{A}_{ex}(t-t_{k-1})}$, $e^{(\mathbf{A}\otimes I_n)(t-t_{k-1})}\to 0$ component-wisely, as $t\to\infty$, so does the left-hand side of (2.28).

It is clear to see that $II_2^T=II_3$. Similar argument can be applied to II_3, which gives

$$\begin{aligned}&e^{\mathbf{A}_{ex}t}\int_{t_{k-1}}^t e^{-\mathbf{A}_{ex}\tau}\text{vec}(II_3)d\tau \\ =&\sum_{j=1}^b\left[\int_{t_{k-1}}^t e^{\mathbf{A}_{ex}(t-\tau)}F_j'e^{\left(I_n\otimes\mathbf{A}^T\right)(\tau-t_{k-1})}d\tau\right]\text{vec}\left(\mathbf{B}_j^T\right) \\ =&\sum_{j=1}^b\text{vec}^{-1}_{n^2\times n^2}\left\{\text{vec}^{-1}_{n^4\times n^4}\left\{\left(P|_{J_M^*}\right)_R^{-1}J_M^{*\,-1}\right.\right. \\ &\left.\left\{P\,\text{vec}\left[I_{n^2}\otimes\left(e^{\mathbf{A}_{ex}(t-t_{k-1})}\mathbf{A}_{ex}^{-1}\right)\right.\right.\right. \\ &\left.\left.\left.-e^{(I_n\times\mathbf{A})(t-t_{k-1})}\otimes\mathbf{A}_{ex}^{-1}\right]\right\}\Big|_{J_M^*}\right\} \\ &\cdot\text{vec}(F_j')\Big\}\,\text{vec}\left(\mathbf{B}_j^T\right)\to 0,\end{aligned} \tag{2.30}$$

as $t\to\infty$, component-wisely and exponentially, where $F_j':=I_n\otimes\left(\bar{\mathbf{F}}_je_X^T\right)$ by Assumption (As), Lemmas A.3 and A.2 with $M=I_{n^8}-I_n\otimes\mathbf{A}^T\otimes I_{n^4}\otimes\mathbf{A}_{ex}^{-1}$ here. The terms P, $P|_{J_M^*}$ and J_M^* in (2.30)

are different from those in (2.29). Finally, let us consider the term containing $\mathrm{vec}(II_1)$ in (2.27):

$$\begin{aligned}
&e^{\mathbf{A}_{ex}t}\int_{t_{k-1}}^{t} e^{-\mathbf{A}_{ex}\tau}\mathrm{vec}(II_1)d\tau\\
&=\sum_{j=1}^{b}\left[\int_{t_{k-1}}^{t} e^{\mathbf{A}_{ex}(t-\tau)}\left(\mathbf{B}_j\otimes\mathbf{B}_j\right)\left(e^{\mathbf{A}}\otimes e^{\mathbf{A}}\right)^{\tau-t_{k-1}}d\tau\right]\mathrm{vec}(e_{XX})\\
&\overset{(A.3)}{=}\sum_{j=1}^{b}\mathrm{vec}^{-1}_{n^2\times n^2}\left\{\mathrm{vec}^{-1}_{n^4\times n^4}\left\{\left(P|_{J_M^*}\right)_R^{-1}J_M^{*\,-1}\right.\right.\\
&\qquad\left\{P\,\mathrm{vec}\left[I_{n^2}\otimes\left(e^{\mathbf{A}_{ex}(t-t_{k-1})}\mathbf{A}_{ex}^{-1}\right)\right.\right.\\
&\qquad\left.-e^{\mathbf{A}^T(t-t_{k-1})}\otimes e^{\mathbf{A}^T(t-t_{k-1})}\right.\\
&\qquad\left.\otimes\mathbf{A}_{ex}^{-1}\right]\Big\}\Big|_{J_M^*}\Big\}\\
&\qquad\left.\cdot\,\mathrm{vec}\left(\mathbf{B}_j\otimes\mathbf{B}_j\right)\right\}\cdot\mathrm{vec}(e_{XX})\to 0
\end{aligned}\tag{2.31}$$

component-wisely and exponentially, as $t\to\infty$, by Lemmas A.3 and A.2 with $M=I_{n^8}-\left(\log(e^{\mathbf{A}^T}\otimes e^{\mathbf{A}^T})\right)^T\otimes I_{n^4}\otimes\mathbf{A}_{ex}^{-1}$, where $e_{XX}:=\bar{\hat{X}}\bar{\hat{X}}^T-\bar{\tilde{X}}\bar{\tilde{X}}^T$, if $\log\left(e^{\mathbf{A}^T}\otimes e^{\mathbf{A}^T}\right)$ exists and unique. In fact, the existence of the logarithm of $\left(e^{\mathbf{A}^T}\otimes e^{\mathbf{A}^T}\right)$ is equivalent to its invertibility (Theorem 1.27, [33]), which is guaranteed by Assumption (As). Its logarithm is unique with all its eigenvalues lying in the strip $\{z\in\mathbb{C}:-\pi\le\mathrm{Im}(z)\le\pi\}$ (Theorem 1.31,[33]), since its eigenvalues are $e^{\lambda_i(\mathbf{A})+\lambda_j(\mathbf{A})}$, $i,j=1,\dots,n$, with real part to be positive, by Assumption (As) again.

Eq. (2.18) follows from (2.23) and (2.27)–(2.31) and the fact that $e_{QP}(t|t_{k-1})=e_{Q\tilde{P}}(t|t_{k-1})-e_{P\tilde{P}}(t|t_{k-1})$. □

Remark 2.2. According to (2.20), if our posterior estimate at t_{k-1} is exactly the conditional expectation, then our estimate coincides with the conditional expectation for all $t\in[t_{k-1},t_k]$.

2.2.3. Proof of (S-3)

Proposition 2.2. *At $t=t_k$, the posterior estimate $\hat{X}(t_k|t_k)$ in the form (2.6) with K_k in (2.7) minimizes* $\mathrm{tr}\,P(t_k|t_k)$.

Proof. First, we derive the evolution equation for $P(t_k|t_k)$:

$$\begin{aligned}
P(t_k|t_k)&=E\left[\left(X(t_k)-\hat{X}(t_k|t_k)\right)\left(X(t_k)-\hat{X}(t_k|t_k)\right)^T\Big|\mathscr{F}_{t_k}\right]\\
&\overset{(2.1),(2.3)}{=}P(t_k|t_{k-1})-K_k\mathbf{C}P(t_k|t_{k-1})-P(t_k|t_{k-1})\mathbf{C}^TK_k^T\\
&\quad+K_k\left[\mathbf{C}P(t_k|t_{k-1})\mathbf{C}^T+\sum_{j=1}^{b}\mathbf{G}_jR_j(t_k)(\mathbf{G}_j)^T\right]K_k^T.
\end{aligned}\tag{2.32}$$

Suppose there exists $\tilde{K}_k\neq 0_{n\times n}$ such that

$$K_k=P(t_k|t_{k-1})\mathbf{C}^T\left[\mathbf{C}P(t_k|t_{k-1})\mathbf{C}^T+\sum_{j=1}^{b}\mathbf{G}_jR_j(t_k)\mathbf{G}_j^T\right]^{-1}+\tilde{K}_k$$

and substituting it into (2.32), we get

$$\begin{aligned}
&P(t_k|t_k)\\
&=P(t_k|t_{k-1})\\
&\quad-P(t_k|t_{k-1})\mathbf{C}^T\left[\mathbf{C}P(t_k|t_{k-1})\mathbf{C}^T+\sum_{j=1}^{b}\mathbf{G}_jR_j(t_k)\mathbf{G}_j^T\right]^{-1}\mathbf{C}P(t_k|t_{k-1})\\
&\quad+\tilde{K}_k\left[\mathbf{C}P(t_k|t_{k-1})\mathbf{C}^T+\sum_{j=1}^{b}\mathbf{G}_jR_j(t_k)\mathbf{G}_j^T\right]\tilde{K}_k^T.
\end{aligned}\tag{2.33}$$

It is clear that the expression in the bracket of the last term on the right-hand side of (2.33) is positive semidefinite, so the trace of the last term never vanishes unless $\tilde{K}_k\equiv 0_{n\times n}$. Therefore, according to (2.33), $P(t_k|t_k)$ is minimized by choosing K_k in (2.7) with $Q(t_k|t_{k-1})$ replaced by $P(t_k|t_{k-1})$. □

Remark 2.3. In our algorithm, $Q(t|t_{k-1})$ takes the place of $P(t|t_{k-1})$. In Proposition 2.1 we show that $Q(t|t_{k-1})$ approaches to $P(t|t_{k-1})$ exponentially fast. Presumably, $Q(t_k|t_{k-1})$ is close to $P(t_k|t_{k-1})$. In Proposition 2.2, we show that $trP(t_k|t_k)$ is minimized by properly chosen K_k. Intuitively, $Q(t_k|t_k)$ is "almost minimized" by choosing K_k in (2.7), since $Q(t_k|t_{k-1})$ is close to $P(t_k|t_{k-1})$.

3. Application to nonlinear filtering problems (NLF)

It is showed in [25] that the nonlinear systems can be approximated by the bilinear system via Carleman approach. This technique will be briefly recalled in Section 3.1. Section 3.2 is devoted to illustrate our algorithm by numerically solving two typical nonlinear examples, i.e., the cubic system with scalar nonlinear observation and three dimensional Lorenz system. The results have been compared with the widely used EKF and UKF.

3.1. Bilinear approximation of nonlinear systems

The continuous–discrete NLF problem considered here is as follows:

$$\begin{cases}
dx_t=\phi(x_t)dt+\sum_{j=1}^{b}F_jdW_j(t)\\
y(t_k)=h(x_{t_k})+\sum_{j=1}^{b}G_jV_j(t_k)
\end{cases},\tag{3.1}$$

where x_t is the state vector in $\mathbb{R}^n$, $y(t_k)$, $k=1,2,\dots,K$ are the discrete measurements in $\mathbb{R}^m$ and $\phi:\mathbb{R}^n\mapsto\mathbb{R}^n$, $h:\mathbb{R}^n\mapsto\mathbb{R}^m$ are smooth nonlinear maps. $V_j(t_k)\sim\mathcal{N}(0,R_j(t_k))$ are independent one dimensional white noises, $W_j(t)$, $j=1,\dots,b$ are independent standard Brownian motion.

Let $(\Omega,\mathscr{F},P)$ be a probability space, $\{\mathscr{F}_{t_k}\}$ with $t_k\in[0,T]$ be a family of nondecreasing σ-algebras of $\mathscr{F}$. Moreover, the initial state x_0 is an $\mathscr{F}_0$-measurable random variable and independent of $W_i(t)$, $i=1,\dots,b$.

Under the assumption of analyticity of ϕ and h, Eq. (3.1) can be rewritten using the Taylor expansion at a given state $\tilde{x}$:

$$\begin{cases}
dx_t=\sum_{i=0}^{\infty}\Phi_i(\tilde{x})(x_t-\tilde{x})^{[i]}dt+\sum_{j=1}^{b}F_jdW_j(t)\\
y(t_k)=\sum_{i=0}^{\infty}H_i(\tilde{x})(x_t-\tilde{x})^{[i]}+\sum_{j=1}^{b}G_jV_j(t_k)
\end{cases}\tag{3.2}$$

with

$$\Phi_i(x)=\frac{1}{i!}\left(\nabla_x^{[i]}\otimes\phi\right),\ H_i(x)=\frac{1}{i!}\left(\nabla_x^{[i]}\otimes h\right),\tag{3.3}$$

where $\star^{[i]}$ is the Kronecker power defined as

$$M^{[0]}=1,\quad M^{[i]}=M\otimes M^{[i-1]}=M^{[i-1]}\otimes M,\tag{3.4}$$

and the operator $\nabla_x^{[i]}\otimes$ applied to a function $\psi=\psi(x):\mathbb{R}^n\to\mathbb{R}^m$ is defined as

$$\nabla_x^{[0]}\otimes\psi=\psi,\ \nabla_x^{[i+1]}\otimes\psi=\nabla_x\otimes\left(\nabla_x^{[i]}\otimes\psi\right),\ i\ge 1\tag{3.5}$$

with $\nabla_x=[\partial/\partial x_1,\dots,\partial/\partial x_n]$.

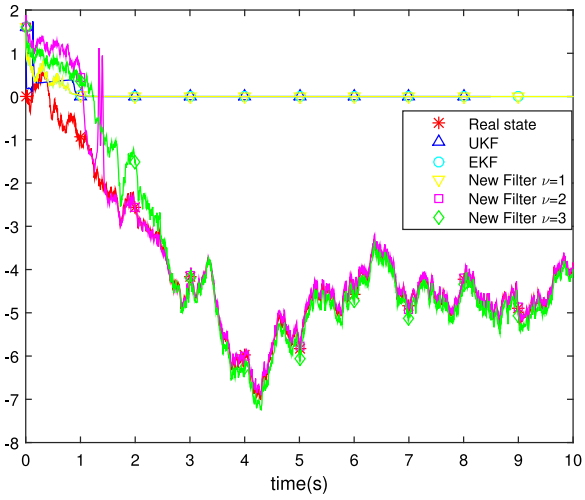


Fig. 1. Estimations of the state in the cubic sensor problem (3.8) via UKF, EKF and SLE with $\nu = 1, 2, 3$.

Given a positive integer ν, let $Y^\nu(t_k) = y(t_k)$ and the original state x_t is augmented as new state

$$X^\nu(t) = \begin{bmatrix} x_t^{[1]} \\ x_t^{[2]} \\ \vdots \\ x_t^{[\nu]} \end{bmatrix} \in \mathbb{R}^{n_\nu}, \ n_\nu = \sum_{i=1}^{\nu} n^i. \tag{3.6}$$

The evolution of the augmented state $X^\nu(t)$ satisfies the following bilinear system.

$$\begin{cases} dX^\nu(t) = \mathbf{A}^\nu(\tilde{x})X^\nu(t)dt + \mathbf{N}^\nu(\tilde{x})dt \\ \qquad\qquad + \sum_{j=1}^{b}(\mathbf{B}_j^\nu X^\nu(t) + \mathbf{F}_j^\nu)dW_j(t) \\ Y^\nu(t_k) = \mathbf{C}^\nu(\tilde{x})X^\nu(t_k) + \mathbf{D}^\nu(\tilde{x}) + \sum_{j=1}^{b}\mathbf{G}_j^\nu V_j(t_k) \end{cases}, \tag{3.7}$$

with $X_0^\nu = \left[x_0^T, \ldots, x_0^{[\nu]T}\right]^T$ and $\mathbf{A}^\nu, \mathbf{N}^\nu, \mathbf{B}_j^\nu, \mathbf{F}_j^\nu, \mathbf{C}^\nu, \mathbf{D}^\nu$ and $\mathbf{G}_j^\nu$ are constant matrices. The specific expressions of these matrices can be found in [25].

Our algorithm will be used to solve (3.7) in the next subsection. The computational complexity of (3.7) is the same as EKF if the augmented state X^ν is with the same size of the state in EKF. If $\nu = 1$, then our algorithm bears the same computational cost as that of EKF. Of course, the higher ν is, the more computational demanding our algorithm is. But one can see from Lorenz system in Section 3.2.2, even with $\nu = 1$ (with the same computational cost as EKF), our algorithm gives more satisfactory estimates, see Fig. 2(c).

Notice that the estimation of the component $x_t^{[1]}$ of $X^\nu(t)$ in (3.6) is the estimate of the original state $x(t)$.

3.2. Simulation

In this subsection, we use two classical examples including the cubic sensor problem and Lorenz system to show the efficiency of our proposed filter and compare the results with those of EKF and UKF.

3.2.1. Cubic sensor

The cubic sensor problem is a benchmark example of essentially infinite-dimensional NLF problem:

$$\begin{cases} dx(t) = dv(t) \\ y(t_k) = x^3(t_k) + w(t_k), \end{cases} \tag{3.8}$$

where $x(t) \in \mathbb{R}$ is the state and $x(t_0) \sim \mathcal{N}(0.2, 1)$, $v \in \mathbb{R}$ is standard Brownian motion, and $y(t_k) \in \mathbb{R}$ is the discrete measurement disturbed by the white noise $w(t_k) \sim \mathcal{N}(0, 1)$.

It can be easily known that the proposed filter with $\nu = 1$ is reduced to the classical EKF since the drift term is $\phi \equiv 0$ in (3.1). Besides, it is widely known that EKF always fails in solving this problem [11]. The approximated bilinear system (3.7) in the cubic sensor problem with $\nu = 2$ is with $X^\nu = \left[x^{[1]}, x^{[2]}\right]^T$ and

$$\begin{aligned} &\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \mathbf{N} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \mathbf{B} \begin{bmatrix} 0 & 0 \\ 2 & 0 \end{bmatrix}, \mathbf{F} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \\ &\mathbf{C} = \left[-3\tilde{x}^2, 3\tilde{x}\right], \mathbf{D} = \tilde{x}^3; \end{aligned} \tag{3.9}$$

while that in the case $\nu = 3$, the corresponding bilinear system (3.7) is with $X^\nu = \left[x^{[1]}, x^{[2]}, x^{[3]}\right]^T$ and

$$\begin{aligned} &\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 3 & 0 & 0 \end{bmatrix}, \mathbf{N} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 0 & 0 & 0 \\ 2 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix}, \\ &\mathbf{F} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \\ &\mathbf{C} = [0, 0, 1], \mathbf{D} = 0. \end{aligned} \tag{3.10}$$

Remark 3.4. It is worthy to notice that the system (3.8) is exactly transformed to the bilinear system (3.10) when $\nu = 3$ without any truncation or approximation.

All ODEs in the simulations are solved by Euler method with initial values $\hat{x}(t_0|t_0) = 0.2$ and $P(t_0|t_0) = 0.1I_\nu$, $I_\nu \in \mathbb{R}^{\nu\times\nu}$ is the identity matrix. The observations are obtained at discrete times $t_k = k\Delta$ with $\Delta = 0.005$ on the interval $[0, T]$ with $T = 10$. The results of different methods for one realization are displayed in Fig. 1. It can be clearly seen that EKF, UKF and SLE with $\nu = 1$ are identical and they all fail completely, while SLE with $\nu = 2,3$ can track the real state quite well. Let us define the mean of the squared estimation error (MSE) for one realization

$$\mu_x = \frac{1}{K+1}\sum_{k=0}^{K}(x_{t_k} - \hat{x}_{t_k})^2, \tag{3.11}$$

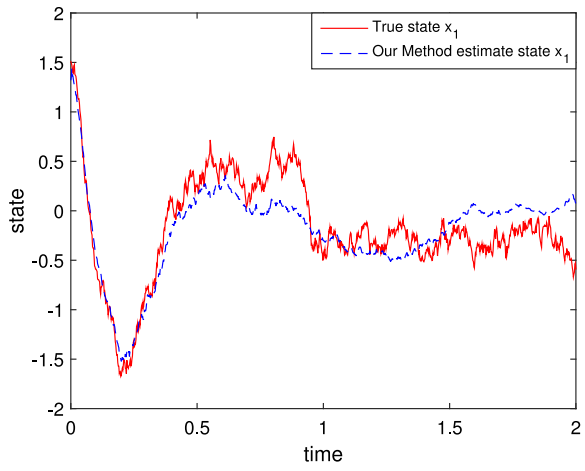
and the MSE averaged over 100 simulations for different methods are listed in Table 1. It is clear that our method outperforms EKF in the average sense and the higher ν yields the better estimate.

3.2.2. Lorenz system

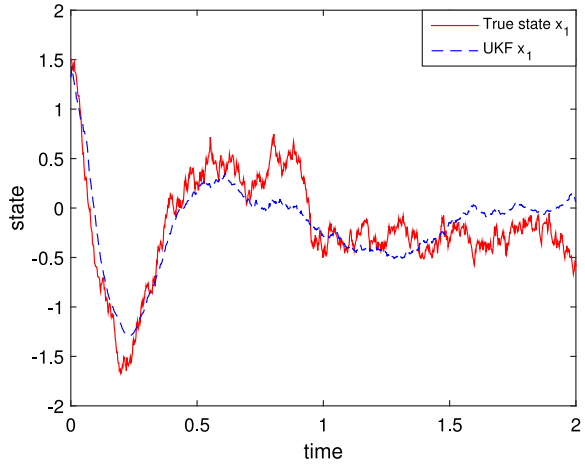
The Lorenz equation considered here is as follows:

$$\begin{cases} dx_1(t) = \sigma(x_2(t) - x_1(t))dt + dv_1(t) \\ dx_2(t) = (\rho x_1(t) - x_2(t) - x_1(t)x_3(t))dt + dv_2(t) \\ dx_3(t) = (x_1(t)x_2(t) - \beta x_3(t))dt + dv_3(t) \\ y(t_k) = [x_1(t_k)x_2(t_k) + x_1(t_k)x_3(t_k) + x_2(t_k)x_3(t_k)]/100 \\ \qquad + w(t_k), \end{cases} \tag{3.12}$$

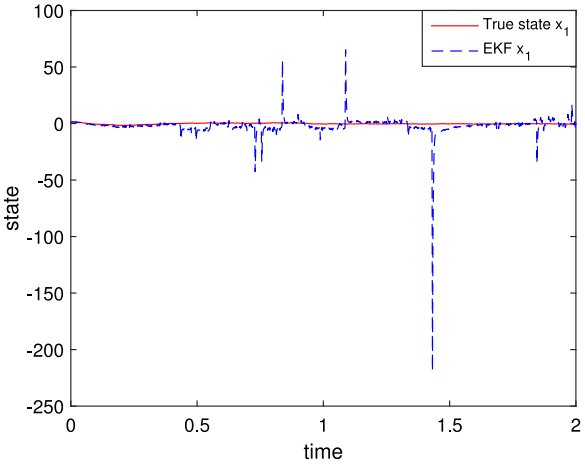
where $\sigma = 5, \beta = 8/3, \rho = -2$ are parameters and $X(t) = [x_1(t), x_2(t), x_3(t)]^T$ is the state of three dimension. The $v_i, i = 1, 2, 3$ are three independent standard Brownian motions. $y(t_k)$ is the discrete measurement where t_k is the instant when measurement arrives and the $w(t_k) \sim \mathcal{N}(0, 0.1^2)$ is the white noise. The numerical realization of system (3.12) is achieved in the



(a) SLE with $\nu = 1$



(b) UKF



(c) EKF

Fig. 2. Estimations of the state x_1 in Lorentz system (3.12) via different methods, including SLE with $\nu = 1$, UKF and EKF.

time interval $[0, 20]$ with initial value $X_0 = [x_{10}, x_{20}, x_{30}]^T = [1.508870, -1.531271, 25.46091]^T$ and time step $\Delta = 0.002$.

Table 1
The average of the MSE over 100 simulations of cubic sensor problem with initial value $\hat{x}(t_0|t_0) = 0.2$ and $P(t_0|t_0) = 0.1I_\nu$.

Algorithm	Average MSE
Our method with $\nu = 1$	–
Our method with $\nu = 2$	2.353321
Our method with $\nu = 3$	0.994113
EKF	–
UKF	–

Table 2
MSE of Lorenz system by our method.

Initial value	MSE of x_1	MSE of x_2	MSE of x_3
$\mathcal{N}(X_0, I_3/10)$	0.2615	0.3298	0.3226

When $\nu = 1$ the new state is $X^\nu(t) = X^{[1]} = X(t) \in \mathbb{R}^3$ and similarly we can get the approximated bilinear system (3.1) with

$$\begin{aligned}
&\mathbf{A} = \begin{bmatrix} -\sigma & \sigma & 0 \\ \rho - \tilde{x}_3 & -1 & -\tilde{x}_1 \\ \tilde{x}_2 & \tilde{x}_1 & -\beta \end{bmatrix}, \mathbf{N} = \begin{bmatrix} 0 \\ \tilde{x}_1\tilde{x}_3 \\ \tilde{x}_1\tilde{x}_2 \end{bmatrix}, \\
&\mathbf{B}_i = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \ i = 1, 2, 3, \\
&\mathbf{F}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \mathbf{F}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \mathbf{F}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \\
&\mathbf{C} = \left[\tilde{x}_2 + \tilde{x}_3, \tilde{x}_1 + \tilde{x}_3, \tilde{x}_1 + \tilde{x}_2\right]/100, \\
&\mathbf{D} = (-\tilde{x}_1\tilde{x}_2 - \tilde{x}_1\tilde{x}_3 - \tilde{x}_2\tilde{x}_3)/100.
\end{aligned} \tag{3.13}$$

It is easy to verify that Assumption (As) in SLE has been satisfied. The initial value for both our method with $\nu = 1$ is $\hat{X}(t_0|t_0) \sim \mathcal{N}(X_0, I_3/10)$. All the ODEs in the simulations are solved by Euler method with step size $\Delta = 0.002$ in the time interval $[0, 2]$. We only display the estimation result of state x_1 in Fig. 2 due to the page limit, and those for x_2, x_3 are similar. Compared the performance of SLE with UKF and EKF, EKF blows up, while SLE can track as well as UKF. The averaged MSE of SLE over 100 simulations are showed in Table 2.

4. Conclusion

In this paper, we construct a SLE for the continuous–discrete bilinear system and apply this algorithm to NLF problems. We show rigorously that the SLE obtained by our algorithm in the prediction is asymptotically approaching to the minimum variance estimate, the conditional expectation, component-wisely and exponentially fast, under the assumption that essentially the system is stable. The update step gives the minimum variance estimate in the linear space spanned by the previous predict estimate and the innovation of observation. The simulations show the efficiency of our method compared to the classical EKF and UKF.

Acknowledgments

Xue Luo acknowledges the support from National Natural Science Foundation of China (grant no. 11501023) and the Fundamental Research Funds for the Central Universities (grant no. YWF-18-BJ-J-238).Stephen S.-T. Yau thanks the financial support of NSFC (grant no. 11471184) and the start-up fund from Tsinghua University.

Appendix

Lemma A.1 derives the formula of integration by parts for Kronecker product.

Lemma A.1. *Suppose A and B are square matrices of size n, then*

$$\int_0^t A(\tau)\otimes(dB(\tau)) = A(\tau)\otimes B(\tau)\Big|_0^t - \int_0^t (dA(\tau))\otimes B(\tau). \tag{A.1}$$

Proof. It is clear that the integration by parts holds for matrix-valued functions, i.e.

$$\int_0^t A(\tau)dB(\tau) = A(\tau)B(\tau)\Big|_0^t - \int_0^t (dA(\tau))B(\tau), \tag{A.2}$$

since for every $i,j=1,\dots,n$, the integration by parts holds for the scalar-valued functions:

$$\begin{aligned}\left(\int_0^t A(\tau)dB(\tau)\right)_{ij} &= \sum_{k=1}^n \int_0^t a_{ik}(\tau)\left(db_{kj}(\tau)\right)\\ &= \sum_{k=1}^n a_{ik}(\tau)b_{kj}(\tau)\Big|_0^t - \sum_{k=1}^n \int_0^t (da_{ik}(\tau))\, b_{kj}(\tau),\end{aligned}$$

where a_{ik} and b_{kj} are the (ik)th element of A and the (kj)th element of B, respectively. Let us look at the (ij)th block of $A\otimes dB$, $i,j=1,\dots,n$:

$$\int_0^t a_{ij}(\tau)(dB(\tau)) \overset{(A.2)}{=} a_{ij}(\tau)B(\tau)\Big|_0^t + \int_0^t da_{ij}(\tau)B(\tau).$$

Thus, (A.1) follows immediately. □

The integral of the matrix-valued function in Lemma A.2 appears frequently in the proof of Proposition 2.1. We compute it here directly using Lemma A.3.

Lemma A.2. *Suppose A and B are square matrices of size n. Assume that the unique* $\log A$ *exists and unique* B^{-1} *exists. Suppose* $M := I_{n^4} - (\log A)^T\otimes I_{n^2}\otimes B^{-1}$ *is similar to the Jordan canonical form* J_M*, i.e.* $M = P^{-1}J_MP$*, with* $\operatorname{rank}(M) = m \le n^2$*. Then we have*

$$\begin{aligned}&\int_0^t A^\tau\otimes e^{B(t-\tau)}d\tau\\ =&\operatorname{vec}_{n^2\times n^2}^{-1}\left\{\left(P|_{J_M^*}\right)_R^{-1}J_M^{*-1}\left\{P\operatorname{vec}\left[I_n\otimes\left(e^{Bt}B^{-1}\right)-A^t\otimes B^{-1}\right]\right\}\Big|_{J_M^*}\right.\\ &\left.+\left[I_{n^2}-\left(P|_{J_M^*}\right)_R^{-1}P|_{J_M^*}\right]w\right\},\end{aligned} \tag{A.3}$$

for arbitrary $w\in\mathbb{C}^{n^2}$*, where* J_M^* *is the square submatrix containing only the non-zero Jordan blocks in* J_M*,* $\circ|_{J_M^*}$ *is the rectangular submatrix of* $\circ$ *keeping the rows which are the non-zeros in* J_M*, and right inverse* $\left(P|_{J_M^*}\right)_R^{-1} := P|_{J_M^*}^T\left(P|_{J_M^*}P|_{J_M^*}^T\right)^{-1}$.

Remark A.1. If M is invertible, then (A.3) is simplified as

$$\begin{aligned}\int_0^t A^\tau\otimes e^{B(t-\tau)}d\tau \;=\; &\operatorname{vec}_{n^2\times n^2}^{-1}\left\{M^{-1}\operatorname{vec}\left[I_n\otimes\left(e^{Bt}B^{-1}\right)\right.\right.\\ &\left.\left.-\;A^t\otimes B^{-1}\right]\right\},\end{aligned} \tag{A.4}$$

where $M := I_{n^4} - (\log A)^T\otimes I_{n^2}\otimes B^{-1}$ as in Lemma A.2.

Proof. For the sake of convenience, let us denote the left-hand side of (A.3) Int for short. If the unique $(\log A)$ exists and B^{-1} exists, then we have

$$\begin{aligned}\text{Int} &= \int_0^t A^\tau\otimes\left[-B^{-1}d\left(e^{B(t-\tau)}\right)\right]\\ &= -\left(I_n\otimes B^{-1}\right)\int_0^t A^\tau\otimes d\left(e^{B(t-\tau)}\right)\\ &\overset{(A.1)}{=} -\left(I_n\otimes B^{-1}\right)\left[A^\tau\otimes e^{B(t-\tau)}\Big|_0^t - \int_0^t d(A^\tau)\otimes e^{B(t-\tau)}\right]\\ &= -\left(I_n\otimes B^{-1}\right)\left[A^t\otimes I_n - I_n\otimes e^{Bt}\right.\\ &\quad\left.-\left(\int_0^t A^\tau\otimes e^{B(t-\tau)}d\tau\right)(\log A\otimes I_n)\right]\\ &= -\left(I_n\otimes B^{-1}\right)\left[A^t\otimes I_n - I_n\otimes e^{Bt}\right] + \left(I_n\otimes B^{-1}\right)\text{Int}\\ &\quad\times(\log A\otimes I_n),\end{aligned} \tag{A.5}$$

where the fourth equality follows from $d(A^\tau) = A^\tau\log A$. Moving the term containing Int to the left-hand side and vectorizing both sides of (A.5), we have

$$\begin{aligned}&\left(I_{n^4} - (\log A)^T\otimes I_{n^2}\otimes B^{-1}\right)\operatorname{vec}(\text{Int})\\ =&\left[I_{n^4} - (\log A\otimes I_n)^T\otimes\left(I_n\otimes B^{-1}\right)\right]\operatorname{vec}(\text{Int})\\ =&\operatorname{vec}\left[I_n\otimes\left(e^{Bt}B^{-1}\right) - A^t\otimes B^{-1}\right].\end{aligned} \tag{A.6}$$

For short, let us denote $M := \left(I_{n^4} - (\log A)^T\otimes I_{n^2}\otimes B^{-1}\right)$ in this proof. If M^{-1} is invertible, (A.4) follows immediately. Otherwise, suppose $\operatorname{rank}(M) = m < n^4$. Without loss of generality, we assume that J_M is the Jordan canonical form of M with the Jordan blocks J_i of size n_i, $i=1,\dots,k$, such that $\sum_{i=1}^k n_i = m < n^4$, i.e. $M = P^{-1}J_MP$. Let us denote J_M^* the square matrix of size m containing only the non-zero Jordan blocks, say the first k diagonal blocks. Then (A.6) becomes

$$\begin{aligned}P|_{J_M^*}\operatorname{vec}(\text{Int}) =&[P\operatorname{vec}(\text{Int})]|_{J_M^*}\\ =&J_M^{*-1}\left\{P\operatorname{vec}\left[I_n\otimes\left(e^{Bt}B^{-1}\right) - A^t\otimes B^{-1}\right]\right\}\Big|_{J_M^*},\end{aligned} \tag{A.7}$$

where $\circ|_{J_M^*}$ is the $m\times n^2$ submatrix of $\circ_{n^2\times n^2}$ with the first m rows, or $\circ|_{J_M^*}$ represents the first m elements, if $\circ$ is a n^2 column vector. It is easy to know that $\operatorname{rank}\left(P|_{J_M^*}\right) = m$, thus there exists a right inverse $\left(P|_{J_M^*}\right)_R^{-1} := P|_{J_M^*}^T\left(P|_{J_M^*}P|_{J_M^*}^T\right)^{-1}$ such that (A.7) has the general solution in the form

$$\begin{aligned}\operatorname{vec}(\text{Int}) =&\left(P|_{J_M^*}\right)_R^{-1}J_M^{*-1}\left\{P\operatorname{vec}\left[I_n\otimes\left(e^{Bt}B^{-1}\right) - A^t\otimes B^{-1}\right]\right\}\Big|_{J_M^*}\\ &+\left[I_{n^2} - \left(P|_{J_M^*}\right)_R^{-1}P|_{J_M^*}\right]w,\end{aligned}$$

for arbitrary $w\in\mathbb{C}^{n^2}$ exists, if and only if

$$P|_{J_M^*}\cdot\left(P|_{J_M^*}\right)_R^{-1} = P|_{J_M^*}P|_{J_M^*}^T\left(P|_{J_M^*}P|_{J_M^*}^T\right)^{-1} = I_m. \quad \square$$

The differences such as $e_X(t|t_{k-1})$, $e_{P\tilde{P}}(t|t_{k-1})$, $e_{Q\tilde{P}}(t|t_{k-1})$, etc. in Proposition 2.1 are claimed to converge to zero component-wisely and exponentially as $t\to\infty$, by repetitively using the following lemma.

Lemma A.3. *If A is a Hurwitz matrix of size n,* $t\in\mathbb{R}^+$*,* $\mathbb{R}^+$ *represents all the positive real numbers, then all elements in the matrix* e^{At} *tends to 0 exponentially fast as* $t\to\infty$*.*

Proof. It is sufficient to show that for all $e_i = (0,\dots,1,\dots,0)^T\in\mathbb{R}^n$, $i=1,\dots,n$, $e^{At}e_i\to 0$ component-wisely and exponentially, as $t\to\infty$, since $\left(e^{At}\right)_{ij} = e_ie^{At}e_j$, $i,j=1,\dots,n$.

It is worth noticing that $x(t) = e^{At}e_i$ is the solution to the ordinary differential equation (ODE) $\dot{x}(t) = Ax(t)$, with the initial value $x(0) = e_i$. Without loss of generality, we assume that A is similar to the Jordan canonical form J, i.e. $A = P^{-1}JP$. Let $\tilde{x}(t) = Px(t)$. It is clear to see that $\tilde{x}(t) = e^{Jt}Pe_i$ is the solution to $\dot{\tilde{x}}(t) = J\tilde{x}(t)$ with the initial condition $\tilde{x}(0) = Pe_i$.

Suppose that there are k Jordan blocks J_j of size n_j, $j=1,\dots,k$. Then the ODE system of $\tilde{x}$ can be decomposed to k decoupled subsystem, i.e. $\dot{\tilde{x}}|_{J_j}(t) = J_j\tilde{x}|_{J_j}(t)$ with the initial value $(Pe_i)|_{J_j}$, where $\circ|_{J_j}$ is the n_j elements corresponding to J_j's rows in J. Due to the

special structure of J_j, $\tilde{x}|_{J_j}(t)$ can be solved explicitly from the n_jth element to the 1st one. It is easy to see that all the elements $\tilde{x}|_{J_j}$ are all of the form $e^{\lambda_j t}p(t)$, where λ_j is the diagonal element of J_j, also one of the eigenvalues of A, and $p(t)$ is a polynomial of t of at most $n_j - 1$ degree. All the elements of $\tilde{x}|_{J_j}$ decays exponentially fast to 0, as $t \to 0$, since A is Hurwitz, i.e. $\mathrm{Re}(\lambda(A)) < 0$. So does $x(t) = P^{-1}\tilde{x}(t) \to 0$. □

References

[1] R.E. Kalman, A new approach to linear filtering and prediction problems, J. Basic Eng. 82 (1) (1960) 35–45.
[2] R.E. Kalman, R.S. Bucy, New results in linear filtering and prediction theory, J. Basic Eng. 83 (1) (1961) 95–108.
[3] A.H. Jazwinski, Stochastic Processes and Filtering Theory, Academic Press, 1970.
[4] R.E. Kalman, New methods in Wiener filtering theory, in: J.L. Bogdanoff, F. Kozin (Eds.), Proceedings of the First Symposium on Engineering Applications of Random Function Theory and Probability, John Wiley & Sons, New York, 1963.
[5] T.E. Duncan, Probability Densities for Diffusion Processes with Applications to Nonlinear Filtering Theory and Detection Theory, Stanford Univ. Ca Stanford Electronics Labs, 1967.
[6] N.E. Mortensen, Optimal Control of Continuous-Time Stochastic Systems, California Univ. Berkeley Electronics Research Lab, 1966.
[7] M. Zakai, On the optimal filtering of diffusion processes, Probab. Theory Related Fields 11 (3) (1969) 230–243.
[8] I. Gyongy, N. Krylov, On the splitting-up method and stochastic partial differential equations, Ann. Probab. 31 (2) (2003) 564–591.
[9] K. Ito, Approximation of the Zakai equation for nonlinear filtering, SIAM J. Control Optim. 34 (2) (1996) 620–634.
[10] N. Nagase, Remarks on nonlinear stochastic partial differential equations: an application of the splitting-up method, SIAM J. Control Optim. 33 (6) (2006) 1716–1730.
[11] X. Luo, S.S.-T. Yau, Complete real time solution of the general nonlinear filtering problem without memory, IEEE Trans. Automat. Control 58 (10) (2013) 2563–2578.
[12] X. Luo, S.S.-T. Yau, Hermite spectral method to 1-D forward Kolmogorov equation and its application to nonlinear filtering problems, IEEE Trans. Automat. Control 58 (10) (2013) 2495–2507.
[13] X. Luo, On recent advance of nonlinear filtering theory: Emphases on global approaches, Pure Appl. Math. Q. 10 (2014) 685–721.
[14] C. Floris, Numeric solution of the Fokker–Planck-Kolmogorov equation, Engineering 5 (12) (2013) 975–988.
[15] S. Sarkka, On unscented Kalman filtering for state estimation of continuous-time nonlinear systems, IEEE Trans. Automat. Control 52 (9) (2007) 1631–1641.
[16] E.A. Wan, R. Van Der Merwe, The unscented Kalman filter for nonlinear estimation, in: Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC, 2000, pp. 153–158.
[17] G. Evensen, The ensemble Kalman filter: Theoretical formulation and practical implementation, Ocean Dyn. 53 (4) (2010) 343–367.
[18] M.S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, IEEE Trans. Signal Process. 50 (2) (2002) 174–188.
[19] A. Gelb, Applied Optimal Estimation, MIT press, 1974.
[20] S. Sarkka, J. Sarmavuori, Gaussian filtering and smoothing for continuous-discrete dynamic systems, Signal Process. 93 (2) (2013) 500–510.
[21] I. Arasaratnam, S. Haykin, T.R. Hurd, Cubature Kalman filtering for continuous-discrete systems: theory and simulations, IEEE Trans. Signal Process. 58 (10) (2010) 4977–4993.
[22] P. Frogerais, J.J. Bellanger, L. Senhadji, Various ways to compute the continuous-discrete extended kalman filter, IEEE Trans. Automat. Control 57 (4) (2012) 1000–1004.
[23] R.R. Mohler, Natural bilinear control processes, IEEE Trans. Syst. Sci. Cybern. 6 (3) (1970) 192–197.
[24] F. Carravetta, A. Germani, M.K. Shuakayev, A new suboptimal approach to the filtering problem for bilinear stochastic differential systems, SIAM J. Control Optim. 38 (4) (2000) 1171–1203.
[25] A. Germani, C. Manes, P. Palumbo, Filtering of differential nonlinear systems via a carleman approximation approach, in: 44th IEEE Conf. on Decision and Control & European Control Conference (CDC-ECC 2005), 2005, pp. 5917–5922.
[26] X. Luo, Y. Jiao, Yang W.-L. Chiou, S.S.-T. Yau, A novel suboptimal method for solving polynomial filtering problems, Automatica 62 (2014) 26–31.
[27] X. Luo, S.S.-T. Yau, The suboptimal nonlinear filtering with augmented states via probabilists' Hermite polynomials, Automatica 94 (2018) 9–17.
[28] F. Cacace, V. Cusimano, A. Germani, P. Palumbo, A state predictor for continuous-time stochastic systems, Systems Control Lett. 98 (2016) 37–43.
[29] F. Cacace, V. Cusimano, A. Germani, P. Palumbo, M. Papi, Optimal linear filter for a class of nonlinear stochastic differential systems with discrete measurements, in: IEEE Conference on Decision and Control, 2017.
[30] F. Carravetta, A. Germani, M. Raimondi, Polynomial filtering for linear discrete time non-Gaussian systems, SIAM J. Control Optim. 34 (5) (1996) 1666–1690.
[31] I. Karatzas, S. Shreve, Brownian Motion and Stochastic Calculus, Springer Science & Business Media, 2012.
[32] R.K. Miller, A.N. Michel, Ordinary Differential Equations, Academic Press, 1982.
[33] N. Higham, Functions of Matrices. Theory and Computation, SIAM, 2008.