System Identification of Nonlinear State-Space Models with Linearly Dependent Unknown Parameters Based on Variational Bayes

Kenji Fujimoto *, Akihiro Taniguchi **, and Yoshiharu Nishida **

Abstract: In this paper, we propose a parameter estimation method for nonlinear state-space models based on the variational Bayes. It is proved that the variational posterior distribution of the hidden states is equivalent to a posterior distribution of the states of an augmented nonlinear state-space model. This enables us to estimate the probability of the hidden states by implementing a variety of existing filtering and smoothing algorithms. Using this technique, a system identification algorithm for nonlinear systems based on variational Bayes and nonlinear smoothers is proposed. It is expected to be more accurate than the existing results since it does not employ any additional approximations in executing the variational Bayes inference. Furthermore, a numerical example demonstrates the effectiveness of the proposed method.

Key Words: Bayesian inference, filtering and smoothing, nonlinear system identification.

1. Introduction

Modern control theory requires a mathematical model of the plant system to be controlled. In particular, for nonlinear control, an accurate state-space model is inevitable. The objective of the present paper is to provide a system identification algorithm for nonlinear systems based on the Bayesian inference. Since a nonlinear state-space model consists of nonlinear functions of constant systems parameters and time varying state variables, such an inference problem reduces to an estimation problem for both unknown constants and unknown hidden states. For this problem, there are some approaches based on maximum likelihood estimation. For instance, Ghahramani and Roweis [1] propose a parameter estimation method based on the EM algorithm, that is, the maximum likelihood estimation for models with hidden variables. However, the maximum likelihood estimation method often has problems with overfitting. Moreover, it is hard to evaluate the reliability of the estimation result. On the other hand, the Bayesian inference, which provides the probability density function of the unknown parameters, prevents the overfitting problem. This is because the inference uses the prior knowledge of the unknown parameters. In addition, it can evaluate the reliability of estimated parameters by the estimated variance. Unfortunately, the integral required for computing the posterior distribution is often analytically intractable, so approximations must be made to obtain it. In such a case, the variational Bayesian inference derives a good approximation of the posterior distribution (variational posterior distribution) by assuming that the unknown parameters and the hidden variables are statistically independent. There exist some attempts to apply the variational Bayesian inference to nonlinear state-space models. Valpola and Karhunen [2] use feedforward neural networks for representing nonlinear functions and approximate the true posterior distribution of the parameters of the networks. However, this method does not treat the input of the model. In addition, the method cannot use any prior knowledge on the nonlinear functions. Daunizeau et al. [3] apply the variational Bayes inference to nonlinear state-space models with inputs. Since nonlinearity of the likelihood function causes the integral in the variational posterior distribution to be analytically intractable, it is approximated by the Laplace approximation. However, the distribution thus obtained may be very different from the true variational posterior distribution since the Laplace approximation only approximates a probability density function near one of the peaks of the true distribution.

This paper proposes a parameter estimation method based on the variational Bayesian inference for nonlinear state-space models that are linear in the unknown parameters. It derives a more accurate variational posterior distribution than the existing results since it does not use any additional approximations such as the Laplace approximation. Instead, we restrict the plant system to be in the class of nonlinear state-space models with linearly dependent unknown parameters which contains a large enough class of practical systems. Furthermore, it is proved that the variational posterior distribution of the hidden states is equivalent to the posterior distribution of the states of an augmented nonlinear state-space model. Therefore, the variational posterior distribution can be obtained by implementing a variety of existing smoothing algorithms. The performance of the proposed algorithm is demonstrated with a numerical example.

Notations used in this paper are defined as follows. For a random variable $x \in \mathbb{R}^n$ with a corresponding probability density function $p(x)$, the expectation of a given function $f(x)$ is denoted by $(f(x))_{\mathbb{E}} := \int f(x)p(x)dx$. The Gaussian distribution with a mean $\mu$ and a covariance matrix $\Sigma$ is denoted by $N(\mu, \Sigma)$. If an $n \times n$ random matrix $\Lambda$ has the Wishart distribution (a multivariate exten-
sion of the chi-squared distribution) with \(r\) degrees of freedom and a scale matrix \(W\), then the distribution is denoted by \(\mathcal{W}(W|v) \coloneqq C_{W}|\Lambda^{(v-\nu-1)/2}\exp[-(1/2)\text{Tr}[W^{-1}\Lambda]]\) where \(C_{W} \coloneqq |W|^{-\nu/2}(2\pi)^{-\nu/2}2^{-\nu/2}/\Gamma(\nu/2)\). Here the scale matrix \(W\) is a symmetric positive definite matrix. The symbol vec(\cdot) denotes a function that maps a matrix to a vector by re-ordering its elements in such a way that vec(A) = \([a_{11}^T, \ldots, a_{n}^T]\) holds for a matrix \(A = [a_1, \ldots, a_n] \in \mathbb{R}^{n \times n}\) whose \(i\)-th column is \(a_i \in \mathbb{R}^m\). The symbol \(\otimes\) denotes the Kronecker product.

2. Variational Bayesian Inference

This section briefly reviews the variational Bayesian inference [4]. In the Bayesian inference [5], the integral required for computing the posterior distribution is often analytically intractable. Therefore, the variational Bayesian inference is proposed to obtain an approximate posterior distribution (variational posterior distribution). In what follows, we consider the case in which there exists two unknown variables the system parameter \(\theta\) and the hidden variable \(X\). Given the measured data \(Y\), an approximation of the true posterior distribution of these variables \(p(X, \theta|Y)\) is denoted by \(q(X, \theta)\).

The goal of the variational Bayesian inference is to derive the approximation of the true posterior distribution which minimizes the following Kullback-Leibler (KL) divergence

\[
\text{KL}[q(X, \theta), p(X, \theta|Y)] = -\int q(X, \theta) \ln \frac{p(X, \theta|Y)}{q(X, \theta)} \text{d}X\text{d}\theta. \tag{1}
\]

The KL divergence \(\text{KL}[q(X, \theta), p(X, \theta|Y)]\) is a measure of the difference between the true posterior distribution \(p(X, \theta|Y)\) and its approximation \(q(X, \theta)\). Hence, minimizing \(\text{KL}[q(X, \theta), p(X, \theta|Y)]\) with respect to \(q(X, \theta)\) gives the best approximation of the true posterior distribution \(p(X, \theta|Y)\).

Assume that the joint distribution \(p(X, Y, \theta)\) is factorized as

\[
p(X, Y, \theta) = p(X, Y|\theta)\prod_{i=1}^{I} p(\theta_i), \tag{2}
\]

where \(\theta = [\theta_1, \ldots, \theta_I]\). To make the minimization of the KL divergence (1) easier, \(q(X, \theta)\) is also factorized as

\[
q(X, \theta) = q(X)\prod_{i=1}^{I} q(\theta_i). \tag{3}
\]

Then, \(q(X)\) and \(q(\theta_i)\) minimizing the KL divergence (1) are derived as follows:

\[
q(X) = C_X \exp(\ln p(X, Y|\theta)|_{q(\theta)}), \tag{4}
\]

\[
q(\theta_i) = C_{\theta_i} p(\theta_i) \exp(\ln p(X, Y|\theta)|_{q(X), q(\theta_{\neq i})})), \tag{5}
\]

where \(\theta_{\neq i}\) denotes \(\{\theta_1, \ldots, \theta_i, \ldots, \theta_I\}\) and, the symbols \(C_X\) and \(C_{\theta_i}\) are the normalizing constants.

Since (4) and (5) cannot be solved analytically, they are computed recursively as follows. Here \(k\) denotes the iteration number and \(q(\theta^{(k)})\) denotes the variational posterior distribution at the \(k\)-th iteration.

Step 2 Compute the following steps until the solution converges.

**VB-E step**

\[
q(X^{(k+1)}) = C_X \exp(\ln p(X, Y|\theta)|_{q(\theta^{(k)})}), \tag{6}
\]

**VB-M step**

For \(i = 1, \ldots, I\)

\[
q(\theta_i) = C_{\theta_i} p(\theta_i) \exp(\ln p(X, Y|\theta)|_{q(X^{(k+1)}), q(\theta_{\neq i})})), \tag{7}
\]

Set \(k \leftarrow k + 1\).

3. Variational Bayes for Nonlinear State-Space Models

In this section, we propose a parameter estimation method based on the variational Bayesian inference for nonlinear state-space models that are linear in unknown parameters.

3.1 Problem Setting

Consider the following nonlinear state-space model:

\[
x_{i+1} = G\phi(x_i, u_i) + v_i, \quad x_i \sim N(x_i, 0, Q),
\]

\[
y_i = H\phi(x_i, u_i) + v_i, \quad v_i \sim N(v_i, 0, R). \tag{8}
\]

Here \(x_i \in \mathbb{R}^m\) is the state, \(u_i \in \mathbb{R}^m\) is the input, \(y_i \in \mathbb{R}^l\) is the output and \(\phi : \mathbb{R}^{m \times m} \to \mathbb{R}^l\) is a known nonlinear function. The matrices \(G \in \mathbb{R}^{m \times m}\) and \(H \in \mathbb{R}^{l \times m}\) consist of the unknown constant elements. The distribution of the initial state \(x_0\) is \(p(x_0) = N(x_0|\mu_0, V_0)\). The sequence of the state \(x_{0:N} = [x_0, \ldots, x_N]\) is the hidden variable and the sequence of the output \(y_{0:N} = [y_0, \ldots, y_N]\) is the measured data. The objective is to derive the variational posterior distribution of the parameters \(G, H, Q\) and \(R\) by applying the variational Bayesian inference to the state-space model (8).

3.2 Prior Distributions

We propose the following candidate prior distributions which will be proved to be conjugate priors [6] in Section 3.4.

\[
p(G|Q) = N(\text{vec}(G)|\mu_G, K \otimes Q), \tag{9}
\]

\[
p(H|R) = N(\text{vec}(H)|\mu_H, M \otimes R), \tag{10}
\]

\[
p(Q) = \mathcal{W}(Q^{-1}|S_Q, \nu), \quad p(R) = \mathcal{W}(R^{-1}|S_R, \eta). \tag{11}
\]

Here \(\mu_G \in \mathbb{R}^m\), \(K \in \mathbb{R}^{m \times m}\), \(S_Q \in \mathbb{R}^{m \times m}\), \(\mu_H \in \mathbb{R}^l\), \(M \in \mathbb{R}^{m \times m}\) and \(S_R \in \mathbb{R}^{l \times l}\) are hyperparameters and \(K, S_Q, M, S_R\) are symmetric positive definite matrices. A conventional method to tune these hyperparameters is to choose the values maximizing prediction performance in a validation set [7].

3.3 Variational Posterior Distribution of the Hidden Variable

In this subsection, we apply the VB-E step to the nonlinear state-space model (8). Barber and Chiappa [8], [9] show that the VB-E step for a linear state-space model corresponds to Kalman smoother of an augmented linear state-space model. Its nonlinear counterpart is proved here. That is, it is shown that the VB-E step for a nonlinear state-space model (8) is equivalent to state estimation using a nonlinear smoother for an augmented nonlinear state-space model. More precisely, the VB-E
step is implemented using a Rauch-Tung-Striebel (RTS) type of smoother for nonlinear state-space models.

The VB-E step for the problem setting in Section 3.1 is

\[
q(x_{0:N}, y_{0:N}^{(k+1)}) \propto \exp \left( \ln p(x_{0:N}, y_{0:N}^{(k)}|\theta) \right) q(\theta),
\]

where \( \theta = [G, H, Q, R] \). Calculating the expectation of the log

\[
\langle \ln p(x_{0:N}, y_{0:N}^{(k)}|\theta) \rangle_{q(\theta)}
\]

where \( \hat{\theta} = [\hat{G}, \hat{H}, \hat{Q}, \hat{R}] \). Hence, for (12), we get \( q(x_{0:N}) = p(x_{0:N}|y_{0:N}^{(k)}, \hat{\theta}) \). The results of the VB-E step required for calculating the 

\[
\text{VB-M step are the variational marginal posterior distributions } q(x_{t}) (t = 0, \ldots, N) \text{ and the covariances of } x_{t} \text{ and } x_{t+1} \text{ denoted by } C_{t+1|t} (t = 0, \ldots, N - 1). \text{ We can calculate them by using a Rauch-Tung-Striebel (RTS) type of smoother for nonlinear state-space models. Therefore, the VB-E step for a nonlinear state-space model (8) coincides with a Rauch-Tung-Striebel (RTS) type of smoother for the augmented nonlinear state-space model (15).}

In this paper, we use the unscented Rauch-Tung-Striebel smoother (URTSS) [10] for the VB-E step. URTSS is the Rauch-Tung-Striebel (RTS) type of smoother for nonlinear state-space models based on the unscented transformation (UT) [11]. The algorithm of URTSS is summarized as follows. Here \( \hat{x}_{t|t} \) denotes the estimation of \( x_{t} \) based on the observations \( y_{1}, \ldots, y_{t} \).

(URTSS)

Step 1

Set the mean and covariance of the initial state \( x_{0} \) as

\[
\hat{x}_{0|0} = \mu_{0}, \quad P_{0|0} = V_{0}.
\]

Step 2

Implement UKF with \( t = 0, \ldots, N. \)

Measurement Update Step

1. Form the set of \( 2n + 1 \) sigma points of the random variable \( x_{t} \) as

\[
X_{ij}^{*} = \begin{cases}
\hat{x}_{t|t-1} \\
(\lambda t_{ij} - 1) / \sqrt{\lambda}
\end{cases} \quad (i = 0, n),
\]

\[
(\lambda t_{ij} + 1) / \sqrt{\lambda}, \quad (i = 1, \ldots, n),
\]

\[
(\lambda t_{ij} + 1) / \sqrt{\lambda}, \quad (i = n + 1, \ldots, 2n),
\]

where \( t_{ij} = \begin{bmatrix} 0 & 1 \end{bmatrix}^T \) is the \( i \)-th column of the matrix \( \sqrt{(n + \lambda)} P_{t|t} \) and \( \lambda \) is a parameter defined by \( \lambda := \kappa(n + \kappa) - n \). How to select the parameters \( \alpha, \beta \) and \( \kappa \) are explained in [12].

2. Transform the sigma points as

\[
Y_{ij} = \hat{H} \phi(X_{ij}^{*}, u_{t}) \quad (i = 0, \ldots, 2n).
\]

3. Compute the following equations

\[
\hat{y}_{t} = \sum_{j=0}^{2n} W_{ij}^{(m)} Y_{ij},
\]

\[
P_{t}^{y} = \sum_{j=0}^{2n} W_{ij}^{(c)} (Y_{ij} - \hat{y}_{t})(Y_{ij} - \hat{y}_{t})^T + R,
\]

\[
P_{t|t} = \sum_{j=0}^{2n} W_{ij}^{(c)} (X_{ij}^{*} - \hat{x}_{t|t-1})(Y_{ij} - \hat{y}_{t})^T,
\]

\[
K_{t} = P_{t|t}^{-1} (P_{t|t}^{y})^{-1},
\]

\[
\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_{t} (Y_{t} - \hat{y}_{t}),
\]

\[
P_{t|t} = P_{t|t-1} - K_{t} P_{t|t-1}^y K_{t}^T.
\]
where \( W_i^{(m)} \) and \( W_i^{(c)} \) are defined as
\[
W_0^{(m)} = \frac{n}{n + \lambda},
\]
\[
W_0^{(c)} = \frac{n}{n + \lambda} + (1 - \alpha^2 + \beta),
\]
\[
W_i^{(m)} = W_i^{(c)} = \frac{1}{2(n + \lambda)} \quad (i = 1, \ldots, 2n).
\]

Prediction Step

1. Form the set of \( 2n + 1 \) sigma points of the random variable \( x_t \) as
\[
X_{t,i} = \begin{cases} 
\hat{x}_{t,i}, & \text{(i = 0,)} \\
\hat{x}_{t,i} + (\sqrt{(n + \lambda)}P_{i|t})_{1}, & \text{(i = 1, \ldots, n,)} \\
\hat{x}_{t,i} - (\sqrt{(n + \lambda)}P_{i|t})_{1}, & \text{(i = n + 1, 2n).}
\end{cases}
\]

2. Transform the sigma points as
\[
\bar{X}_{t+1,i} = G\phi(X_{t,i}, u_t) \quad (i = 0, \ldots, 2n).
\]

3. Compute the following equations.
\[
\bar{x}_{t+1,i} = \sum_{i=0}^{2n} W_i^{(m)} X_{t+1,i},
\]
\[
\bar{P}_{t+1,i} = \sum_{i=0}^{2n} W_i^{(c)} (X_{t+1,i} - \bar{x}_{t+1,i}) (X_{t+1,i} - \bar{x}_{t+1,i})^T + Q_t.
\]
\[
\bar{C}_{t+1,i} = \sum_{i=0}^{2n} W_i^{(c)} (X_{t+1,i} - \hat{x}_{t+1,i}) (X_{t+1,i} - \hat{x}_{t+1,i})^T.
\]

Step 3

Calculate the smoother gain \( D_t \), the smoothed mean \( \hat{x}_{t|N} \) and the smoothed covariance \( P_{t|N} \) with \( t = N, \ldots, 0 \) as
\[
D_t = \frac{P_{t+1,i}}{P_{t+1,i} + 1},
\]
\[
\hat{x}_{t|N} = \hat{x}_{t+i} + D_t(\hat{x}_{t+1|N} - \hat{x}_{t+1|i}),
\]
\[
P_{t|N} = P_{t+1|N} + D_t(P_{t+1|N} - P_{t+1|i})D_t^T.
\]

The smoothed cross-covariance \( C_{t+1|N} \) (\( t = 0, \ldots, N - 1 \)) are computed in the VB-M step and can also be calculated by
\[
C_{t+1|N} = D_t P_{t+1|N}.
\]

3.4 Variational Posterior Distribution of the Parameters

Suppose that the variational posterior distribution of the parameters \( q(G, Q, H, R) \) is factorized as
\[
q(G, Q, H, R) = q(G, Q)q(H, R).
\]

Then, the VB-M step for the problem setting in Section 3.1 is
\[
q(G, Q) = \exp \left( \ln p(x_0, x_N, y_0) \right) q(\mu_G, \hat{K} \otimes Q).
\]
\[
q(H, R) = \exp \left( \ln p(x_0, x_N, y_0) \right) q(\mu_H, \hat{M} \otimes R).
\]

Calculating (41) and (42) using the prior distributions (9)–(11), we get
\[
q(G|Q) = N(\mu_G, \hat{K} \otimes Q),
\]
\[
q(H|R) = N(\mu_H, \hat{M} \otimes R),
\]
\[
q(R) = \exp (\ln p(x_0, x_N, y_0) q(\mu_R, \hat{\Sigma} \otimes R).
\]

Consequently, the proposed algorithm is summarized as follows. Here \( \hat{X}_{t|N} \) and \( P_{t|N} \) are the mean and the covariance of \( x_t \) with the corresponding probability density function \( q(x_t) \).

(Proposed Algorithm)

Step 1
Set the initial distributions \( q(G|Q) \), \( q(H|R) \) and \( q(R) \) as in (9)–(11) and \( k \leftarrow 0 \).

Step 2
Compute the following steps until the solution converges.

VB-E Step
Calculate the marginal distributions \( q(x_t)^{(k+1)} \) (\( t = 0, \ldots, N \)) and covariances \( C_{t+1|N}^{(k+1)} \) (\( t = 0, \ldots, N-1 \)) by using URTSS (19)–(38) and (39) as
\[
q(x_t) = N(\hat{x}_{t|N}, P_{t|N}).
\]

VB-M Step
Calculate the following variational posterior distributions by using (43)–(46) and Appendix as
\[
q(G|Q) = N(\mu_G, \hat{K} \otimes Q).
\]
\[
q(H|R) = N(\mu_H, \hat{M} \otimes R),
\]
\[
q(R) = \exp (\ln p(x_0, x_N, y_0) q(\mu_R, \hat{\Sigma} \otimes R).
\]

Set \( k \leftarrow k + 1 \).

4. Evaluation of the Proposed Algorithm

In this section, the performance of the proposed algorithm is demonstrated using a numerical simulation.

4.1 Example

Consider the following system
\[
x_{t+1} = \begin{bmatrix} x_{1,t} + 0.01 x_{2,t} \\ -x_{1,t} - 0.1 x_{2,t} + 0.1 u_t + 3 x_{1,t} \\ w_t \end{bmatrix} + w_t
\]
\[
y_t = 10 x_{1,t} + v_t
\]
\[
= H \phi(x_t, u_t) + v_t,
\]
\[
w_t \sim N(0, Q), \quad v_t \sim N(0, R),
\]

where \( w_t \) and \( v_t \) are independent white noises.
where
\[ G = \begin{bmatrix} 1 & 0.01 & 0 & 0 \\ -1 & -0.1 & 0.1 & 3 \end{bmatrix}, \quad H = [10, 0, 0, 0], \]
\[ \phi(x_t, u_t) = [x_t^T, u_t, x_{1,t}^3]^T, \]
\[ Q = 0.00001I_2, \quad R = 0.01. \]

The symbol \( x_{1,t} \) denotes the first element of \( x_t \in \mathbb{R}^2 \). The signal to noise ratio of the external noise is \(-20\)dB. The initial state is \( x_0 = 0 \). The time interval is \( N = 250 \). The hyperparameters of the prior distributions (9)–(11) and the distribution of \( x_0 \) are given as follows.
\[ \mu_G = \begin{bmatrix} 0.1, & 0.1, & 0.1, & 0.1, & 0.1, & 0.1 \end{bmatrix}^T, \]
\[ \mu_H = \begin{bmatrix} 0.1, & 0.1, & 0.1, & 0.1 \end{bmatrix}^T, \]
\[ K = 10I_2, \quad M = 10I_2, \quad \mu_0 = 0, \quad V_0 = 10I_2, \]
\[ S_Q = I_2, \quad \nu = 1, \quad S_K = 1, \quad \eta = 1. \]
The input is a scalar \( u_t \sim N(0, 1000) \). The parameters of UT are \( \alpha = 0.001, \beta = 2 \) and \( k = 0 \).

4.2 Results

When the time step is \( k = 500 \), the means of the variational posterior distributions \( q(G|Q), q(Q), q(H|R) \) and \( q(R) \) are as follows.
\[ G_{\text{est}} = \begin{bmatrix} -0.0142 & 0.183 & 0.0205 & 0.0105 \\ 0.4654 & 0.9016 & -0.0014 & 0.0117 \end{bmatrix}, \]
\[ H_{\text{est}} = \begin{bmatrix} 0.0366 & 1.0656 & -0.0005 & 0.0083 \end{bmatrix}, \]
\[ Q_{\text{est}} = \begin{bmatrix} 0.0425 & -0.006 \\ -0.006 & 0.0208 \end{bmatrix}, \quad R_{\text{est}} = 0.0192. \]

Figure 1 shows the comparison between the output of the true system \( y \) and its estimation \( y_{\text{est}} \), which is computed using an estimated model \( y_{\text{est}} = G_{\text{est}}\phi(x_t, u_t) + w_t, \quad y_{\text{est}2} = H_{\text{est}}\phi(x_t, u_t), \) a sampled input signal \( u_t \sim N(0, 1000) \) and the initial state is \( x_0 = 0 \). Note that the output signal of the true system used in the comparison is different from that used in estimating the system parameters \( G, H, Q \) and \( R \). The fitting rate of the data shown in Fig. 1 is 43.1\%. The fitting rate is defined by
\[ \text{(fitting rate)} = \left( 1 - \frac{\sum_{t=0}^{N}(y_t - y_{\text{est}, t})^2}{\sum_{t=0}^{N}(y_t - \bar{y})^2} \right) \times 100\% \quad (63) \]
where \( \bar{y} := (1/N) \sum_{t=0}^{N} y_t \). It confirms the accuracy of the estimated output. Figure 2 shows the history of the maximum eigenvalue of the covariance matrix \( \hat{K} \otimes Q_{\text{est}} \) along the iteration number \( k \). A smaller value of the maximum eigenvalue of \( \hat{K} \otimes Q_{\text{est}} \) means a higher reliability of the estimated parameter \( G_{\text{est}} \).

4.3 Discussion

If a value of the nonlinear term (the fourth term of \(-x_{1,t} - 0.1x_{2,t} + 0.1u_t + 3x_{1,t}^3 \) in (52)) is too small in a training set, we cannot properly estimate the parameter of the nonlinear term. Therefore, we use as a training set the data where the nonlinear term is not negligible. In the training set, the state variable \( x_{1,t} \) generally ranges from \(-0.2 \) to \( 0.4 \) (see Fig. 3). When \( x_{1,t} = 0.2 \), the first term \(-x_{1,t} \) is \(-0.2 \) whereas the fourth term \( 3x_{1,t}^3 \) is \( 0.024 \), that is, about 1/10 of the first term. Thus, the value of the nonlinear term is modest.

The estimated parameter \( G_{\text{est}} \) and \( H_{\text{est}} \) are different from their true values \( G \) and \( H \), respectively. This is because there exist in general multiple state-space models describing the same input-output map. In fact, Fig. 1 and the fitting rate indicate that the estimated model well captures the input-output relationship of the true system.

Figure 2 shows that the maximum eigenvalue of the covariance matrix \( \hat{K} \otimes Q_{\text{est}} \) decreases, which means that the estimation \( G_{\text{est}} \) becomes more accurate, as the iteration number \( k \) increases. This result shows that the proposed method estimates a nonlinear state-space model with its reliability accurately. We obtain a similar result for \( \hat{M} \otimes R_{\text{est}} \) as well.

Fig. 2 History of the maximum eigenvalue of the estimated covariance matrix \( \hat{K} \otimes Q_{\text{est}} \) along the iteration number \( k \).
5. Conclusion

This paper proposes a variational Bayesian inference method for nonlinear state-space models that are linear in the unknown parameters. It is proved that the VB-E step for the nonlinear state-space model coincides with a Rauch-Tung-Striebel (RTS) type of smoother for the augmented nonlinear state-space model. Using this analysis, a concrete identification procedure based on variational Bayes is proposed. Furthermore, the numerical example demonstrates the effectiveness of the proposed algorithm.

References

[1] Z. Ghahramani and S.T. Roweis: Learning nonlinear dynamical systems using an EM algorithm, Advances in Neural Information Processing Systems, Vol. 11, pp. 599–605, 1999.

[2] H. Valpola and J. Karhunen: An unsupervised learning method for nonlinear dynamic state-space models, Neural Computation, Vol. 14, No. 11, pp. 2647–2692, 2002.

[3] J. Daunizeau, K.J. Friston, and S.J. Kiebel: Variational Bayes inference and prediction of stochastic nonlinear systems using an EM algorithm, Physica D: Nonlinear Phenomena, Vol. 238, No. 21, pp. 2089–2118, 2009.

[4] H. Attias: Inferring parameters and structure of latent variable models by variational Bayes, Machine Learning, Vol. 20, No. 3, pp. 145–160, 1995.

[5] G.E.P. Box: Sampling and Bayes’ inference in scientific modeling and robustness, Journal of the Royal Statistical Society, Series A (General), Vol. 143, No. 4, pp. 383–430, 1980.

[6] D. Persi and Y. Donald: Conjugate priors for exponential families, The Annals of Statistics, Vol. 7, No. 2, pp. 269–281, 1979.

[7] C.M. Bishop: Pattern Recognition and Machine Learning, sec. 1.3, Springer, 2006.

[8] D. Barber and S. Chiappa: Unified inference for variational Bayesian linear Gaussian state-space models, Proceedings of 20th Conference on Neural Information Processing Systems, pp. 1–8, 2007.

[9] K. Fujimoto, A. Satoh, and S. Fukunaga: System identification based on variational Bayes method and the invariance under coordinate transformations, Proceedings of IEEE Conference on Decision and Control, pp. 3882–3888, 2011.

[10] S. Särkkä: Unscented Rauch–Tung–Striebel smoother, IEEE Transactions on Automatic Control, Vol. 53, No. 3, pp. 845–849, 2008.

[11] S.J. Julier and J.K. Uhlmann: Unscented filtering and nonlinear estimation, Proceedings of the IEEE, Vol. 92, No. 3, pp. 401–422, 2004.

[12] E.A. Wan and R. van der Merwe: The unscented Kalman filter, Kalman Filtering and Neural Networks, pp. 221–280, Wiley, 2001.

Appendix Update Rule of the Variational Posterior Distribution of Parameters

This appendix gives the detailed computation to update the parameters in the algorithm proposed in Section 3.4. Using $\text{mat}(\cdot)$ denoting the inverse of $\text{vec}(\cdot)$, the updated hyperparameters in (43)–(46) are as follows:

\[
\begin{align*}
\mathbf{K}^{-1} &:= \mathbf{K}^{-1} + \mathbf{G}, \\
\hat{\mu}_G &:= \mathbf{K}^{-1} + \mathbf{G}^{-1} \mathbf{K}^{-1} \mathbf{G}^{-1} @ I, \mu_G \\
\hat{S}_G^{-1} &:= \mathbf{S}_G^{-1} + \mathbf{W}_G + \mathbf{mat}(\mu_G) \mathbf{K}^{-1} \times (\mathbf{I} - (\mathbf{K}^{-1} + \mathbf{G}^{-1} \mathbf{K}^{-1}) \mathbf{mat}(\mu_G))\mathbf{Y} \\
&- \mathbf{S}_G^{-1}(\mathbf{K}^{-1} + \mathbf{G}^{-1}) \mathbf{K}^{-1} \mathbf{mat}(\mu_G)^T \\
&- \mathbf{S}_G^{-1} \mathbf{G}^{-1} (\mathbf{K}^{-1} + \mathbf{G}^{-1}) \mathbf{K}^{-1} \mathbf{mat}(\mu_G)^T, \\
\end{align*}
\]

\[
\begin{align*}
\hat{\mu}_H &:= \{\mathbf{M}^{-1} + \mathbf{W}_H\}^{-1} \mathbf{M}^{-1} \mathbf{I} \mu_H \\
&+ \{\mathbf{M}^{-1} + \mathbf{W}_H\}^{-1} \mathbf{I} \mathbf{vec}(\mathbf{S}_H), \\
\hat{S}_H^{-1} &:= \mathbf{S}_H^{-1} + \mathbf{Y}_H + \mathbf{mat}(\mu_H) \mathbf{M}^{-1} \times (\mathbf{I} - (\mathbf{M}^{-1} + \mathbf{W}_H)^{-1} \mathbf{M}^{-1}) \mathbf{mat}(\mu_H)^T \\
&- \mathbf{S}_H^{-1} (\mathbf{M}^{-1} + \mathbf{W}_H)^{-1} \mathbf{M}^{-1} \mathbf{mat}(\mu_H)^T \\
&- \mathbf{S}_H^{-1} (\mathbf{M}^{-1} + \mathbf{W}_H)^{-1} \mathbf{M}^{-1} \mathbf{mat}(\mu_H)^T, \\
\end{align*}
\]

where $\mathbf{W}_G$, $\mathbf{S}_G$, $\mathbf{Y}_H$, $\mathbf{S}_H$ and $\mathbf{W}_H$ are defined by

\[
\begin{align*}
\mathbf{W}_G := \sum_{t=0}^{N-1} \mathbf{x}_{t+1} \mathbf{G}^{-1} \mathbf{x}_{t+1}^T \\
\mathbf{S}_G := \sum_{t=0}^{N-1} \phi(x_t, u_t) \mathbf{G}^{-1} \mathbf{x}_{t+1} \mathbf{G}^{-1} \mathbf{x}_{t+1}^T \\
\mathbf{V}_G := \sum_{t=0}^{N-1} \phi(x_t, u_t) \mathbf{G}^{-1} \mathbf{x}_{t+1} \mathbf{G}^{-1} \mathbf{x}_{t+1}^T \\
\mathbf{Y}_H := \sum_{t=0}^{N-1} \mathbf{y}_t \mathbf{G}^{-1} \mathbf{x}_{t+1} \mathbf{G}^{-1} \mathbf{x}_{t+1}^T \\
\mathbf{W}_H := \sum_{t=0}^{N-1} \phi(x_t, u_t) \mathbf{G}^{-1} \mathbf{x}_{t+1} \mathbf{G}^{-1} \mathbf{x}_{t+1}^T.
\end{align*}
\]

In this paper, we calculate $\mathbf{W}_G$, $\mathbf{S}_G$, $\mathbf{V}_G$, $\mathbf{Y}_H$, $\mathbf{S}_H$ and $\mathbf{W}_H$ approximately by using UT [11]. UT is a method to calculate approximately the mean and the variance of a random variable which undergoes a nonlinear transformation if the mean and the covariance of the random variable before transformation are given. See [11] for the detail of UT. First, we describe how to calculate $\mathbf{W}_G$, $\mathbf{S}_G$ and $\mathbf{V}_G$. Consider propagating a random variable $\mathbf{x}_t$, $\mathbf{x}_{t+1}$ through a nonlinear function $z_t = g(x_t, x_{t+1}) = \{\phi(x, u)\}T$. Note that $u_t$ is given. The mean and the covariance of the random variable $\mathbf{x}_t$, $\mathbf{x}_{t+1}$ are as follows.

\[
\begin{align*}
\hat{\mu}_{t|N} &:= \mathbf{x}_t^T \mathbf{x}_{t+1|N}^T, \\
\hat{P}_{t|N} &:= \mathbf{C}_{t+1|N} \mathbf{C}_{t+1|N}^T, \\
\end{align*}
\]

Since the mean $\hat{\mu}_{t|N}$ and the covariance $\hat{P}_{t|N}$ are known, we can calculate the mean and the covariance of the random variable $z_t$ using UT. Using the mean and the covariance of $z_t$, we obtain

\[
\begin{align*}
\mathbf{V}_G := \sum_{t=0}^{N-1} \langle z_t \rangle \\
\mathbf{S}_G := \sum_{t=0}^{N-1} \langle z_t \rangle \langle z_t \rangle^T + \sum_{t=0}^{N-1} \mathbf{P}_r, \\
\end{align*}
\]
respectively. Then, in the same manner as the calculation of $W_G$, $S_G$ and $V_G$, we can calculate the mean and the covariance of $X_t$ and obtain

\[
S_H = \sum_{t=0}^{N} \langle X_t \rangle q(x_t) x_t^T, \quad (A.14)
\]

\[
W_H = \sqrt{\sum_{t=0}^{N} \langle X_t X_t^T \rangle}, \quad (A.15)
\]

\[
= \sum_{t=0}^{N} \langle X_t \rangle \langle X_t \rangle^T + \sum_{t=0}^{N} P_{X_t}, \quad (A.16)
\]

where $\langle X_t \rangle$ and $P_{X_t}$ are the mean and the covariance of $X_t$, respectively.

---

**Kenji Fujimoto** (Member)

He received his B.S. and M.S. degrees in Engineering and Ph.D. degree in Informatics from Kyoto University, Japan, in 1994, 1996 and 2001, respectively. He is currently a professor of Graduate School of Engineering, Kyoto University, Japan. From 1997 to 2004 he was a research associate of Graduate School of Engineering and Graduate School of Informatics, Kyoto University, Japan. From 2004 to 2012 he was an associate professor of Graduate School of Engineering, Nagoya University, Japan. His research interests include non-linear control and stochastic systems theory.

**Akihiro Taniguchi**

He received his B.S. and M.S. degrees in Engineering from Kyoto University, Japan, in 2015 and 2017, respectively. Since then he has worked in Kobe Steel, Ltd. His research interests include Bayesian inference and process control.

**Yoshiharu Nishida**

He received his B.S., M.S., and Ph.D. degrees in engineering from Osaka Prefecture University, Osaka, Japan, in 1987, 1989, and 1999, respectively. Since 1989, he has been working in Kobe Steel, Ltd., Kobe, Japan. His research interests include control and identification methods for robot, construction machine, mechanical and process control systems. He is a member of ISIJ and ISCIE.