2018 European Control Conference (ECC)
June 12-15, 2018. Limassol, Cyprus

On System Identification of Nonlinear State-Space Models Based on Variational Bayes: Multimodal Distribution Case

Akihiro Taniguchi[1], Kenji Fujimoto[2] and Yoshiharu Nishida[1]

Abstract— **In this paper, we propose a parameter estimation method for nonlinear state-space models based on the variational Bayes. It is shown that the variational posterior distribution of the hidden states is equivalent the probability estimated by a nonlinear smoother of an augmented nonlinear state-space model. This enables us to obtain the variational posterior distribution of the hidden states by implementing a variety of existing nonlinear filtering and smoothing algorithms. By employing a Gaussian mixture distribution as a candidate probability density function of the hidden states, we propose an algorithm to compute multimodal posterior distributions which are not able to be handled by the existing results.**

I. INTRODUCTION

In modern control theory, which is based on state-space models, we need to know a state-space model of a given plant system for controller design. Furthermore, since hidden states in the state-space model are also unknown, we need to infer both the model parameters and the hidden states. There exists a method to estimate them with an extended Kalman filter by treating model parameters as extra hidden states [1]. However, it can not estimate the covariances the unknown parameters and that of the external noise of the state-space model. On the other hand, Roweis and Ghahramani [2] proposed a parameter estimation method for nonlinear state-space models based on the EM algorithm, that is, the maximum likelihood estimation method for models with hidden variables. It can estimate not only the model parameters but also the covariances of the external noise. However, the maximum likelihood estimation method often has problems with overfitting. Moreover, it is hard to evaluate the reliability of the estimated parameters. On the other hand, the Bayesian inference, which provides the probability density functions 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 using their variances. Unfortunately, the posterior distribution is obtained by an integral which is not analytically computable, so approximations must be employed to obtain it. In such a case, the variational Bayesian inference can effectively estimates the approximate posterior distribution (variational posterior distribution) by assuming that the unknown parameters and the hidden variables are statistically independent. There exist some applications of the variational Bayesian inference to nonlinear state-space models. For instance, the paper [3] uses feedforward neural networks for representing nonlinear functions in the model and approximate the true posterior distribution. However, it does not cope with the input of the model. In addition, it cannot use the prior knowledge of the nonlinear functions. On the other hand, [4] applies the variational Bayes inference to nonlinear state-space models with inputs. Since nonlinearity in the likelihood function cause the integral in computing the variational posterior distribution to be analytically intractable, they derive the variational posterior distribution approximately by using the Laplace approximation. However, since the Laplace approximation approximates a probability density function near one of the peaks of the true distribution, this approach can not estimate multimodal probability distributions. The multimodality of the probability function is an important characteristic of the hidden states of nonlinear state-space systems, since it is often produced by the nonlinearity of the state equations.

This paper proposes a parameter estimation method based on the variational Bayesian inference for nonlinear state-space models that are affine in the unknown parameters. It derives a variational posterior distribution without using any unimodal approximations such as the Laplace approximation. In particular, we prove that the posterior probability of the hidden states are equivalent to that derived by a nonlinear smoother for of an augmented nonlinear state-space model. This allows us to use a variety of existing nonlinear smoothers and filters to compute the estimated probability distributions of both the states and the parameters. In addition, by employing a Gaussian mixture distribution as a candidate probability density function of the hidden states, we propose an algorithm to compute multimodal posterior distributions which are not able to be handled by the existing results.

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 $\langle f(x) \rangle_{p(x)} := \int f(x)p(x)\mathrm{d}x$. The Gaussian distribution is denoted by $\mathcal{N}(x|\mu, \Sigma) := (2\pi)^{-n/2}\|\Sigma\|^{-1/2}\exp\{-(1/2)(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\}$. If the distribution of a scalar random variable $x \in \mathbb{R}$ is the gamma distribution, then it is denoted by $\mathrm{Gam}(x \mid \gamma, \lambda) := \frac{\lambda^\gamma}{\Gamma(\gamma)}x^{\gamma-1}\mathrm{e}^{-\lambda x} \quad (x \geq 0)$.

*This work was not supported by any organization

[1] Akihiro Taniguchi and Yoshiharu Nishida are with Kobe Steel, Ltd., 5-5, Takatsukadai 1-chome, Nishi-ku, Kobe, Hyogo, 651-2271, Japan taniguchi.akihiro@kobelco.com, nishida.yoshiharu@kobelco.com

[2] Kenji Fujimoto is with Department of Aeronautics and Astronautics, Kyoto University, Nishikyo-ku, Kyoto, 615-8540, Japan k.fujimoto@ieee.org

II. Variational Bayesian Inference

This section briefly reviews the variational Bayesian inference [5]. In the Bayesian inference [7], the integral required for deriving the posterior distribution is often analytically intractable. Therefore, the variational Bayesian inference is proposed as a way 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 θ 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

$$\begin{aligned}&\mathrm{KL}[q(X,\theta), p(X,\theta|Y)]\\ &:= -\int q(X,\theta)\log\frac{p(X,\theta|Y)}{q(X,\theta)}\mathrm{d}X\mathrm{d}\theta. \end{aligned} \tag{1}$$

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

$$q(X) = C_X \exp\langle \log p(X,Y|\theta)\rangle_{q(\theta)} \tag{4}$$

$$q(\theta_i) = C_{\theta_i} p(\theta_i)\exp\langle \log p(X,Y|\theta)\rangle_{q(X),q(\theta_{-i})} \tag{5}$$

where θ_{-i} is defined by $\theta_{-i} := \{\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_I\}$ and C_X and C_{θ_i} are normalizing constants.

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

(Variational Bayesian Inference)
Step 1 Set the initial distribution $q(\theta)^{(0)}$ and $k \leftarrow 0$.
Step 2 Compute the following steps until the solution converges.
VB-E step

$$q(X)^{(k+1)} = C_X \exp\langle \log p(X,Y|\theta)\rangle_{q(\theta)^{(k)}} \tag{6}$$

VB-M step
For $i = 1, \ldots, I$

$$q(\theta_i) = C_{\theta_i} p(\theta_i)\exp\langle \log p(X,Y|\theta)\rangle_{q(X)^{(k+1)},q(\theta_{-i})^{(k)}} \tag{7}$$

Set $k \leftarrow k+1$.

III. 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 affine in unknown parameters.

A. Problem setting

Consider the following nonlinear state-space model

$$x_{t+1} = f(x_t, u_t, \theta) + w_t, \tag{8}$$

$$y_t = g(x_t, u_t, \phi) + v_t \tag{9}$$

where $x_t \in \mathbb{R}^n$ is a state, $u_t \in \mathbb{R}^m$ is an input, $y_t \in \mathbb{R}^l$ is an output, and $\theta \in \mathbb{R}^{\kappa_1}$ and $\phi \in \mathbb{R}^{\kappa_2}$ are unknown parameters, respectively. The functions f and g are decomposed as

$$f_i(x_t, u_t, \theta) = \theta^{\mathrm{T}} f_{i(1)}(x_t, u_t) + f_{i(0)}(x_t, u_t) \tag{10}$$

$$g_j(x_t, u_t, \phi) = \phi^{\mathrm{T}} g_{j(1)}(x_t, u_t) + g_{j(0)}(x_t, u_t). \tag{11}$$

The symbol $w_t \sim \mathcal{N}(0, \alpha I_n)$ is a Gaussian process noise, and $v_t \sim \mathcal{N}(0, \beta I_l)$ is a Gaussian measurement noise. The distribution of the initial state x_0 is

$$p(x_0) = \mathcal{N}(x_0|\mu_0, V_0). \tag{12}$$

The sequence of the state $X_N = \{x_0, \ldots, x_N\}$ is regarded as a set of hidden states, and the sequence of the input $U_N = \{u_0, \ldots, u_N\}$ and that of the output $Y_N = \{y_0, \ldots, y_N\}$ are observed data. The objective is to derive the variational posterior distributions of the parameters $\theta, \phi, \alpha, \beta$ and the hidden states X_N by applying the variational Bayesian inference to the state-space model (8) and (9). In what follows, we assume that the variational posterior distribution $q(\theta, \phi, \alpha, \beta, X_N)$ is factorized as

$$q(\theta,\phi,\alpha,\beta,X_N) = q(\theta,\alpha)q(\phi,\beta)q(X_N). \tag{13}$$

B. Prior distributions

We propose the following prior distributions which will be proved to be conjugate priors in the next subsection.

$$p(\theta,\alpha) = \mathcal{N}(\theta \mid \mu_\theta, \alpha K)\mathrm{Gam}(\alpha^{-1} \mid \gamma_1, \lambda_1) \tag{14}$$

$$p(\phi,\beta) = \mathcal{N}(\phi \mid \mu_\phi, \beta M)\mathrm{Gam}(\beta^{-1} \mid \gamma_2, \lambda_2). \tag{15}$$

Here $K \in \mathbb{R}^{\kappa_1 \times \kappa_1}$ and $M \in \mathbb{R}^{\kappa_2 \times \kappa_2}$ are symmetric positive definite matrices.

C. Variational posterior distribution of the parameters

In this subsection, we apply VB-M Step to the nonlinear state-space model (8) and (9). VB-M Step for the problem setting in Subsection III-A is

$$q(\theta,\alpha) \propto p(\theta,\alpha) \exp\left\langle \ln p(X_N, Y_N \mid \theta,\phi,\alpha,\beta,U_N)\right\rangle_{q(X_N),q(\phi,\beta)} \tag{16}$$

$$q(\phi,\beta) \propto p(\phi,\beta) \exp\left\langle \ln p(X_N, Y_N \mid \theta,\phi,\alpha,\beta,U_N)\right\rangle_{q(X_N),q(\theta,\alpha)}. \tag{17}$$

Calculating (16) and (17) using the prior distributions (14) and (15), we get

$$q(\theta,\alpha) = \mathcal{N}(\theta \mid \hat{\mu}_\theta, \alpha\hat{K})\mathrm{Gam}(\alpha^{-1} \mid \hat{\gamma}_1, \hat{\lambda}_1) \tag{18}$$

$$q(\phi,\beta) = \mathcal{N}(\phi \mid \hat{\mu}_\phi, \beta\hat{M})\mathrm{Gam}(\beta^{-1} \mid \hat{\gamma}_2, \hat{\lambda}_2), \tag{19}$$

where $\hat{\mu}_\theta$, $\hat{K}$, $\hat{\gamma}_1$, $\hat{\lambda}_1$, $\hat{\mu}_\phi$, $\hat{M}$, $\hat{\gamma}_2$ and $\hat{\lambda}_2$ are the updated hyper parameters of μ_θ, K, γ_1, λ_1, μ_ϕ, M, γ_2 and λ_2, respectively. See Appendix A for the detail of the computation.

D. Variational posterior distribution of the hidden states

This subsection applies VB-E Step to the nonlinear state-space model (8) and (9). Barber and Chiappa [6] show that VB-E Step for a linear state-space model corresponds to a Kalman smoother of an augmented linear state-space model. Its nonlinear counter part is derived here, that is, it is proved that VB-E Step for a nonlinear state-space model (8) and (9) corresponds to a smoothing problem for an augmented nonlinear state-space model.

VB-E Step for the problem setting in Subsection III-A is

$$q(X_N) = C_x \exp\left\langle \ln p(X_N, Y_N|\Theta, U_N)\right\rangle_{q(\Theta)}, \quad (\Theta = \{\theta,\phi,\alpha,\beta\}) \tag{20}$$

where C_x is a constant normalizing the distribution $q(X_N)$. The complete data log likelihood in (20) is

$$\begin{aligned}
&\ln p(X_N, Y_N \mid \Theta, U_N)\\
&= -\frac{1}{2}(x_0-\mu_0)^{\mathrm{T}} V_0^{-1}(x_0-\mu_0)\\
&-\frac{1}{2}\sum_{t=0}^{N-1}\left(x_{t+1}-f(x_t,u_t,\theta)\right)^{\mathrm{T}}\alpha^{-1}I\\
&\left(x_{t+1}-f(x_t,u_t,\theta)\right)\\
&-\frac{1}{2}\sum_{t=0}^{N}\left(y_t-g(x_t,u_t,\phi)\right)^{\mathrm{T}}\beta^{-1}I\left(y_t-g(x_t,u_t,\phi)\right)\\
&+\mathrm{const.}
\end{aligned} \tag{21}$$

Here we treat the terms which do not depend on the hidden states as constant values. Calculating the expectation of the complete data log likelihood, we then obtain

$$\begin{aligned}
&\left\langle \ln p(X_N, Y_N|\Theta, U_N)\right\rangle_{q(\Theta)}\\
&= -\frac{1}{2}(x_0-\mu_0)^{\mathrm{T}} V_0^{-1}(x_0-\mu_0)\\
&-\frac{1}{2}\sum_{t=0}^{N-1}\left\langle\alpha^{-1}\right\rangle\left(x_{t+1}-f(x_t,u_t,\langle\theta\rangle)\right)^{\mathrm{T}}\\
&\left(x_{t+1}-f(x_t,u_t,\langle\theta\rangle)\right)\\
&-\frac{1}{2}\sum_{t=0}^{N}\left\langle\beta^{-1}\right\rangle\left(y_t-g(x_t,u_t,\langle\phi\rangle)\right)^{\mathrm{T}}\\
&\left(y_t-g(x_t,u_t,\langle\phi\rangle)\right)\\
&-\frac{1}{2}\sum_{t=0}^{N}\left\{\sum_{i=1}^{n} f_{i(1)}(x_t,u_t)^{\mathrm{T}}\hat{K}f_{i(1)}(x_t,u_t)\right.\\
&\left.+\sum_{j=1}^{l} g_{j(1)}(x_t,u_t)^{\mathrm{T}}\hat{M}g_{j(1)}(x_t,u_t)\right\}\\
&+\mathrm{const.}
\end{aligned} \tag{22}$$

Let us extend the output y_t, the nonlinear function g and the covariance matrix of the measurement noise $R = \beta I_l$ to

$$\tilde{y}_t := \begin{bmatrix} y_t \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{bmatrix}, \quad \tilde{g}(x_t,u_t,\langle\phi\rangle) := \begin{bmatrix} g(x_t,u_t,\langle\phi\rangle) \\ L^{\mathrm{T}} f_{1(1)}(x_t,u_t) \\ \vdots \\ L^{\mathrm{T}} f_{n(1)}(x_t,u_t) \\ W^{\mathrm{T}} g_{1(1)}(x_t,u_t) \\ \vdots \\ W^{\mathrm{T}} g_{l(1)}(x_t,u_t) \end{bmatrix},$$

$$\tilde{R} := \begin{bmatrix} \left\langle\beta^{-1}\right\rangle^{-1} I_l & O \\ O & I_{n\kappa_1+l\kappa_2} \end{bmatrix}, \tag{23}$$

where the matrices L and W are derived by using the Cholesky factorization of $\hat{K}$ and $\hat{M}$ as $\hat{K} = LL^{\mathrm{T}}$ and $\hat{M} = WW^{\mathrm{T}}$, respectively. Using the extended output, nonlinear function and covariance matrix in (23) and rearranging the expectation of the complete data log likelihood (22), we finally obtain

$$\begin{aligned}
&\left\langle \ln p(X_N, Y_N|\Theta, U_N)\right\rangle_{q(\Theta)} =\\
&-\frac{1}{2}(x_0-\mu_0)^{\mathrm{T}} V_0^{-1}(x_0-\mu_0)\\
&-\frac{1}{2}\sum_{t=0}^{N-1}\left(x_{t+1}-f(x_t,u_t,\langle\theta\rangle)\right)^{\mathrm{T}}\langle\alpha^{-1}\rangle I\\
&\left(x_{t+1}-f(x_t,u_t,\langle\theta\rangle)\right)\\
&-\frac{1}{2}\sum_{t=0}^{N}\left(\tilde{y}_t-\tilde{g}(x_t,u_t,\langle\phi\rangle)\right)^{\mathrm{T}}\tilde{R}^{-1}\left(\tilde{y}_t-\tilde{g}(x_t,u_t,\langle\phi\rangle)\right)\\
&+\mathrm{const.}
\end{aligned} \tag{24}$$

The comparison of (24) with (21) shows that the expectation of the complete data log likelihood corresponds to the

complete data log likelihood of the following augmented state-space model

$$x_{t+1} = f(x_t, u_t, \langle\theta\rangle) + \tilde{w}_t, \tag{25}$$

$$\tilde{y}_t = \tilde{g}(x_t, u_t, \langle\phi\rangle) + \tilde{v}_t, \tag{26}$$

where the distributions of the process noise $\tilde{w}_t$ and the measurement noise $\tilde{v}_t$ are

$$p(\tilde{w}_t) = \mathcal{N}(\tilde{w}_t|0, \langle\alpha^{-1}\rangle^{-1} I),\ p(\tilde{v}_t) \quad = \mathcal{N}(\tilde{v}_t|0, \tilde{R}). \tag{27}$$

Therefore, we have

$$\langle \ln p(X_N, Y_N|\Theta, U_N)\rangle_{q(\Theta)} = \ln \tilde{p}(X_N, \tilde{Y}_N|U_N) + \text{const.}, \tag{28}$$

where $\ln \tilde{p}(X_N, \tilde{Y}_N|U_N)$ denotes the complete data log likelihood of the augmented model (25) and (26). We substitute (28) into (20) and then obtain

$$q(X_N) = \tilde{p}(X_N \mid \tilde{Y}_N, U_N). \tag{29}$$

The above equation shows that the variational posterior distribution of the hidden states of the original state-space model (8) and (9) is equivalent to the probability distribution of the state estimated by a nonlinear smoother for the augmented model (25) and (26). Hence we can obtain the variational posterior distribution by implementing a variety of existing filtering and smoothing algorithms. The proposed method can express the multimodality of the variational posterior distribution by constructing a filtering and smoothing algorithm which calculates a multimodal posterior distribution.

There are the following filtering and smoothing algorithms to calculate the multimodal posterior distribution of a nonlinear state-space model. Kitagawa [10] proposes a stochastic method which calculates a filtered density and smoothed density based on Monte Carlo simulation. In recent years, Havlak and Campbell [11] develop a filtering method for nonlinear state-space models using Gaussian mixture models. Lee and Campbell [12] present a Gaussian sum smoother for the models. This smoother is the nonlinear version of the closed-form Gaussian sum smoother [13] and based on the linearization of the nonlinear functions of the models. Applying these methods to VB-E step enables us to derive the multimodal variational posterior distribution of hidden states.

In this paper, we use the filtering algorithm [11] and the smoothing algorithm [12] to obtain the variational posterior distribution. We explain these algorithms briefly in the rest of this subsection. Now we consider the following nonlinear state-space model without inputs for simplicity

$$x_{t+1} = f(x_t) + w_t, \quad w_t \sim \mathcal{N}(0, Q), \tag{30}$$

$$y_t = g(x_t) + v_t, \quad v_t \sim \mathcal{N}(0, R) \tag{31}$$

and assume that the filtered density at time t is the following Gaussian mixture

$$p(x_t \mid Y_t) = \sum_{i=1}^{N_t} w_t^{(i)} \mathcal{N}(x_t \mid m_t^{(i)}, P_t^{(i)}). \tag{32}$$

We briefly explain the filter [11]. Under the assumption (32), we derive the filtered density at time $t+1$ in the following steps. First, in the same way as the unscented transformation (UT) [9], form the set of $2n+1$ sigma points from each mixand $\mathcal{N}(x_t \mid m_t^{(i)}, P_t^{(i)})$ and transform the sigma points as $\chi_{t+1}^j = f(\chi_t^j) \quad (j = 0, \ldots, 2n)$. We then calculate the linearity criteria defined as

$$e_{\text{res}} := \min_{A,b}\Big(\Big\|\chi_{t+1} - (A\chi_t + b)\Big\|\Big) \quad (A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n), \tag{33}$$

where $\chi_t = [\chi_t^0, \ldots, \chi_t^{2n}]$ and $\|\cdot\|$ denotes the Frobenius norm. This criteria shows how well the optimal linear model approximates the nonlinear transformation of sigma points. If e_{res} is smaller than a threshold, we calculate the predictive density for the mixand in the same way as Unscented Kalman Filter (UKF) [9]. If e_{res} is larger than the threshold, we split the mixand into the Gaussian mixture whose mixand has a smaller variance than the mixand before the splitting has. We then calculate the predictive densities for the mixands of the Gaussian mixture in the same fashion as UKF. With the above procedures, we finally obtain predictive density $p(x_{t+1}|Y_t)$ as

$$p(x_{t+1}|Y_t) = \sum_{i=1}^{\overline{N}_{t+1}} \overline{w}_{t+1}^{(i)} \mathcal{N}(x_{t+1} \mid \overline{m}_{t+1}^{(i)}, \overline{P}_{t+1}^{(i)}). \tag{34}$$

The filtered density $p(x_{t+1} \mid Y_{t+1})$ is derived in a similar way to the predictive step. We briefly explain the Gaussian sum smoother [12]. We consider the state-space model without inputs (30) and (31). The smoothed density $p(x_t \mid Y_N)$ $(t \leq N)$ can be factorized as

$$p(x_t \mid Y_N) = \int p(x_t, x_{t+1} \mid Y_N) \mathrm{d}x_{t+1} \tag{35}$$

$$= p(x_t \mid Y_t) B_{t|N}(x_t) \tag{36}$$

$$B_{t|N}(x_t) := \begin{cases} 1 & (t = N) \\ \int \frac{p(x_{t+1}|Y_N)}{p(x_{t+1}|Y_t)} p(x_{t+1}|x_t) \mathrm{d}x_{t+1} & (t \leq N-1) \end{cases}$$

where $B_{t|N}(x_t)$ is called a backward corrector (BC). If the functions f and g are linear and the filtered density is a Gaussian mixture, then the smoothed density is derived as a Gaussian mixture from (36) [13]. If they are nonlinear, the smoothed density is obtained approximately as the following Gaussian mixture by the Jacobian linearization of the nonlinear system and then adopting the above result

$$p(x_t \mid Y_N) \approx \frac{1}{r_t} \sum_{i=1}^{N_t} \tilde{w}_t^{(i)} \mathcal{N}(x_t \mid \tilde{m}_t^{(i)}, \tilde{P}_t^{(i)}), \tag{37}$$

where the definitions of $\tilde{w}_t^{(i)}$, $\tilde{m}_t^{(i)}$ and $\tilde{P}_t^{(i)}$ are in [12].

As shown in Appendix, we need not only the variational posterior distribution of a hidden state $q(x_t)$ but also the variational posterior distribution of a joint state $q(x_t, x_{t+1})$ in order to calculate VB-M step. For this reason, we extend the smoother [12] so that we can obtain the smoothed density

of the joint state. We can describe BC in (36) in the following recursive form [13]

$$B_{t|N}(x_t) = \int B_{t+1|N}(x_{t+1}) L_t(y_{t+1}; x_{t+1}) p(x_{t+1}|x_t) \mathrm{d}x_{t+1} \tag{38}$$

where

$$L_t(y_{t+1}; x_{t+1}) := \frac{p(y_{t+1}|x_{t+1})}{\int p(y_{t+1}|x_{t+1}) p(x_{t+1}|Y_t) \mathrm{d}x_{t+1}}. \tag{39}$$

In what follows, ν_{t+1} denotes the denominator of (39). Substituting (32), (38) and (39) into (36), we obtain

$$\begin{aligned} & p(x_t, x_{t+1} \mid Y_N) && (40)\\ =& p(x_t|Y_t) B_{t+1|N}(x_{t+1}) L_t(y_{t+1}; x_{t+1}) p(x_{t+1}|x_t) && (41)\\ \approx& \frac{1}{r_{t+1}\nu_{t+1}} \sum_{i=1}^{N_t} w_t^{(i)} \mathcal{N}(x_t \mid m_t^{(i)}, P_t^{(i)}) \times \\ & \mathcal{N}\left(y_{t+2:N} \mid \zeta_{t+2:N}(x_{t+1}), D_{t+2:N}(x_{t+1})\right) \times \\ & \mathcal{N}(y_{t+1} \mid g(x_{t+1}), R) \mathcal{N}(x_{t+1} \mid f(x_t), Q). && (42) \end{aligned}$$

Linearizing the nonlinear functions of x_{t+1} in (42) around $x_{t+1} = f(x_t)$, we then have

$$\begin{aligned} & p(x_t, x_{t+1} \mid Y_N) && (43)\\ =& \frac{1}{r_{t+1}\nu_{t+1}} \sum_{i=1}^{N_t} w_t^{(i)} \mathcal{N}(x_t \mid m_t^{(i)}, P_t^{(i)}) \times \\ & \mathcal{N}\left(y_{t+1:N} \mid \zeta_{t+1:N}(x_t), D_{t+1:N}(x_t)\right) \times \\ & \mathcal{N}\left(x_{t+1} \mid m_{t+1}(x_t), P_{t+1}(x_t)\right), && (44) \end{aligned}$$

where

$$\tilde{C}_{t+2:N}(x_t) := \begin{bmatrix} C_{t+2:N}(f(x_t)) \\ \frac{\partial g}{\partial x_{t+1}}\big|_{x_{t+1}=f(x_t)} \end{bmatrix}, \tag{45}$$

$$m_{t+1}(x_t) := f(x_t) + K_{t+1}(x_t)\left(y_{t+1:N} - \zeta_{t+1:N}(x_t)\right), \tag{46}$$

$$P_{t+1}(x_t) := \left(I_n - K_{t+1}(x_t)\tilde{C}_{t+2:N}(x_t)\right) Q, \tag{47}$$

$$K_{t+1}(x_t) := Q\tilde{C}_{t+2:N}(x_t)^{\mathrm{T}} D_{t+1:N}(x_t)^{-1} \tag{48}$$

and the definitions of the symbols C, D and ζ are given in [12]. In the same way as we get (37), we obtain

$$\begin{aligned} & p(x_t, x_{t+1} \mid Y_N) \approx \\ & \frac{1}{r_{t+1}\nu_{t+1}} \sum_{i=1}^{N_t} \tilde{w}_t^{(i)} \mathcal{N}(x_t \mid \tilde{m}_t^{(i)}, \tilde{P}_t^{(i)}) \\ & \times \mathcal{N}\left(x_{t+1} \mid m_{t+1}(x_t), P_{t+1}(x_t)\right) && (49) \end{aligned}$$

We then express the joint smoothed distribution $p(x_t, x_{t+1} \mid Y_N)$ as a Gaussian mixture model in the following way. First, we approximate the covariance $P_{t+1}(x_t)$ as $P_{t+1}(x_t) \approx P_{t+1}(\tilde{m}_t^{(i)})$. Second, we calculate the mean $\tilde{m}_{t+1}^{(i)}$, covariance $\tilde{P}_{t+1}^{(i)}$ and cross-covariance $\tilde{C}_{t+1}^{(i)}$ of the hidden state x_{t+1} by using UT. This procedure enables us to obtain the joint smoothed density $p(x_t, x_{t+1} \mid Y_N)$ as the following Gaussian mixture model

$$\begin{aligned} & p(x_t, x_{t+1} \mid Y_N) \approx \frac{1}{r_{t+1}\nu_{t+1}} \sum_{i=1}^{N_t} \tilde{w}_t^{(i)} \times \\ & \mathcal{N}\left(\begin{bmatrix} x_t \\ x_{t+1} \end{bmatrix} \middle| \begin{bmatrix} \tilde{m}_t^{(i)} \\ \tilde{m}_{t+1}^{(i)} \end{bmatrix}, \begin{bmatrix} \tilde{P}_t^{(i)} & \tilde{C}_{t+1}^{(i)} \\ \left(\tilde{C}_{t+1}^{(i)}\right)^{\mathrm{T}} & \tilde{P}_{t+1}^{(i)} \end{bmatrix} \right) && (50) \end{aligned}$$

E. Proposed Algorithm

By summarizing the previous subsections, we propose the following algorithm to estimate the unknown parameters of nonlinear state-space models based on variational Bayes.

(Proposed Method)
Step 1
Set the initial distributions $q(\theta, \alpha)^{(0)}$ and $q(\phi, \beta)^{(0)}$ as in (14)–(15) and $k \leftarrow 0$.
Step 2
Compute the following steps until the solution converges.
VB-E Step
Calculate the following distributions by filtering and smoothing algorithms based on [11] and [12].

$$q(x_t)^{(k+1)} \approx \frac{1}{r_t} \sum_{i=1}^{N_t} \tilde{w}_t^{(i)} \mathcal{N}(x_t \mid \tilde{m}_t^{(i)}, \tilde{P}_t^{(i)}) \tag{51}$$

$$\begin{aligned} q(x_t, x_{t+1})^{(k+1)} \approx & \frac{1}{r_{t+1}\nu_{t+1}} \sum_{i=1}^{N_t} \tilde{w}_t^{(i)} \times \\ & \mathcal{N}\left(\begin{bmatrix} x_t \\ x_{t+1} \end{bmatrix} \middle| \begin{bmatrix} \tilde{m}_t^{(i)} \\ \tilde{m}_{t+1}^{(i)} \end{bmatrix}, \begin{bmatrix} \tilde{P}_t^{(i)} & \tilde{C}_{t+1}^{(i)} \\ \left(\tilde{C}_{t+1}^{(i)}\right)^{\mathrm{T}} & \tilde{P}_{t+1}^{(i)} \end{bmatrix} \right) \end{aligned} \tag{52}$$

VB-M Step
Calculate the following variational posterior distributions by using (18)–(19).

$$q(\theta, \alpha)^{(k+1)} = \mathcal{N}(\theta \mid \hat{\mu}_\theta, \alpha\hat{K}) \mathrm{Gam}(\alpha^{-1} \mid \hat{\gamma}_1, \hat{\lambda}_1) \tag{53}$$

$$q(\phi, \beta)^{(k+1)} = \mathcal{N}(\phi \mid \hat{\mu}_\phi, \beta\hat{M}) \mathrm{Gam}(\beta^{-1} \mid \hat{\gamma}_2, \hat{\lambda}_2) \tag{54}$$

Set $k \leftarrow k + 1$.

IV. Conclusion

This paper proposes a variational Bayesian inference method for nonlinear state-space models that are affine in the unknown parameters. It is shown that the variational posterior distribution of the hidden states is equivalent the probability estimated by a nonlinear smoother of an augmented nonlinear state-space model. This enables us to obtain the variational posterior distribution of the hidden states by implementing a variety of existing nonlinear filtering and smoothing algorithms. In particular, by employing a nonlinear filter and a smoother for Gaussian mixture distributions, multimodal probability distributions of the hidden states can be estimated.

REFERENCES

[1] Ljung, L., & Söderström, T.(1983). *Theory and Practice of Recursive Identification*. MIT Press.

[2] Ghahramani, Z., & Roweis, S. T. (1999). Learning nonlinear dynamical systems using an EM algorithm. *Advances in Neural Information Processing Systems*, 11, 599–605.

[3] Valpola, H., & Karhunen, J.(2002). An unsupervised learning method for nonlinear dynamic state-space models. *Neural Computation*, 14(11), 2647–2692.

[4] Daunizeau, J., Friston, K. J., & Kiebel, S. J.(2009). Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models. *Physica D: Nonlinear Phenomena*, 238(21), 2089–2118.

[5] Attias, H.(1999). Inferring parameters and structure of latent variable models by variational Bayes. In *Proc. 15th Conf. on Uncertainty in Artificial Intelligence*, Stockholm, Sweden (pp.21–30).

[6] Barber, D., & Chiappa, S.(2007). Unified inference for variational Bayesian linear Gaussian state-space models. *Advances in Neural Information Processing Systems*, 19, 81–88.

[7] Box, G. E. P.(1980). Sampling and Bayes' Inference in Scientific Modeling and Robustness. *Journal of the Royal Statistical Society*, 143(4), 383–430.

[8] Persi, D., & Donald, Y.(1979). Conjugate priors for exponential families. *The Annals of Statistics*, 7(2), 269–281.

[9] Julier, S. J., & Uhlmann, J. K.(2004). Unscented Filtering and Nonlinear Estimation. In *Proc. IEEE*, 92(3), 401–422.

[10] Genshiro, K.(1996). Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models. *Journal of Computational and Graphical Statistics*, 5(1), 1–25.

[11] Havlak, F., & Campbell, M.E.(2013). Discrete and Continuous, Probabilistic Anticipation for Autonomous Robots in Urban Environments. *arXiv*.

[12] Lee, D. J. & Campbell, M.E.(2015). Smoothing Algorithm for Nonlinear Systems Using Gaussian Mixture Models. *Journal of Guidance, Control, and Dynamics*, 38(8), 1438–1451.

[13] Vo, B. N., Vo, B.T.,& Mahler, R.P.S.(2012). Closed-Form Solutions to Forward-Backward Smoothing. *IEEE Transactions on Signal Processing*, 60(1), 2–17.

APPENDIX

UPDATE RULE OF THE VARIATIONAL POSTERIOR DISTRIBUTION OF PARAMETERS

The updated hyperparameters $\hat{K}$, $\hat{\mu}_\theta$, $\hat{\gamma}_1$ and $\hat{\lambda}_1$ in (18) are

$$\hat{K} := \left(K^{-1} + E_f\right)^{-1}, \tag{55}$$

$$\hat{\mu}_\theta := \hat{K}\left(K^{-1}\mu_\theta + d_f\right), \tag{56}$$

$$\hat{\gamma}_1 := \gamma_1 + \frac{nN}{2}, \tag{57}$$

$$\hat{\lambda}_1 := \lambda_1 + \frac{1}{2}(a_f - 2b_f + c_f + \mu_\theta^{\mathrm{T}} K^{-1}\mu_\theta - \hat{\mu}_\theta^{\mathrm{T}}\hat{K}^{-1}\hat{\mu}_\theta), \tag{58}$$

where a_f, b_f, c_f, d_f and E_f are

$$a_f := \sum_{t=0}^{N-1} \mathrm{Tr}\left[\langle x_{t+1}\rangle_{q(x_{t+1})}\langle x_{t+1}\rangle_{q(x_{t+1})}^{\mathrm{T}} + \mathrm{Var}[x_{t+1}]\right], \tag{59}$$

$$b_f := \sum_{t=0}^{N-1} \left\langle x_{t+1}^{\mathrm{T}} f_0(x_t, u_t)\right\rangle_{q(x_{t+1}, x_t)}, \tag{60}$$

$$c_f := \sum_{t=0}^{N-1} \left\langle f_0(x_t, u_t)^{\mathrm{T}} f_0(x_t, u_t)\right\rangle_{q(x_t)}, \tag{61}$$

$$d_f := \sum_{t=0}^{N-1} \langle h_f(x_{t+1}, x_t)\rangle_{q(x_{t+1}, x_t)}, \tag{62}$$

$$E_f := \sum_{t=0}^{N-1} \langle F(x_t, u_t)\rangle_{q(x_t)}. \tag{63}$$

Here $f_0(x_t, u_t)$, $h_f(x_{t+1}, x_t, u_t)$ and $F(x_t, u_t)$ are defined as

$$f_0(x_t, u_t) := \begin{bmatrix} f_{1(0)}(x_t, u_t) \\ \vdots \\ f_{n(0)}(x_t, u_t) \end{bmatrix}, \tag{64}$$

$$h_f(x_{t+1}, x_t, u_t) := \sum_{i=1}^{n} \left\{(x_{t+1})_i - f_{i(0)}(x_t, u_t)\right\} f_{i(1)}(x_t, u_t), \tag{65}$$

$$F(x_t, u_t) := \sum_{i=1}^{n} f_{i(1)}(x_t, u_t) f_{i(1)}(x_t, u_t)^{\mathrm{T}}. \tag{66}$$

Note that the expectations in a_f, b_f, c_f, d_f and E_f can be calculated by existing methods such as UT. Similarly, the updated hyperparameters $\hat{M}$, $\hat{\mu}_\phi$, $\hat{\gamma}_2$ and $\hat{\lambda}_2$ in (19) are

$$\hat{M} := \left(M^{-1} + E_g\right)^{-1}, \tag{67}$$

$$\hat{\mu}_\phi := \hat{M}\left(M^{-1}\mu_\phi + d_g\right), \tag{68}$$

$$\hat{\gamma}_2 := \gamma_2 + \frac{l(N+1)}{2}, \tag{69}$$

$$\hat{\lambda}_2 := \lambda_2 + \frac{1}{2}\left(a_g - 2b_g + c_g + \mu_\phi^{\mathrm{T}} M^{-1}\mu_\phi - \hat{\mu}_\phi^{\mathrm{T}}\hat{M}^{-1}\hat{\mu}_\phi\right), \tag{70}$$

where a_g, b_g, c_g, d_g and E_g are

$$a_g := \sum_{t=0}^{N} y_t^{\mathrm{T}} y_t, \tag{71}$$

$$b_g := \sum_{t=0}^{N} y_t^{\mathrm{T}} \langle g_0(x_t, u_t)\rangle_{q(x_t)}, \tag{72}$$

$$c_g := \sum_{t=0}^{N} \left\langle g_0(x_t, u_t)^{\mathrm{T}} g_0(x_t, u_t)\right\rangle_{q(x_t)}, \tag{73}$$

$$d_g := \sum_{t=0}^{N} \langle h_g(x_t, u_t)\rangle_{q(x_t)}, \tag{74}$$

$$E_g := \sum_{t=0}^{N} \langle G(x_t, u_t)\rangle_{q(x_t)}. \tag{75}$$

Here $g_0(x_t, u_t)$, $h_g(x_t, u_t)$ and $G(x_t, u_t)$ are defined as

$$g_0(x_t, u_t) := \begin{bmatrix} g_{1(0)}(x_t, u_t) \\ \vdots \\ g_{l(0)}(x_t, u_t) \end{bmatrix}, \tag{76}$$

$$h_g(x_t, u_t) := \sum_{j=1}^{l} \left\{(y_t)_j - g_{j(0)}(x_t, u_t)\right\} g_{j(1)}(x_t, u_t), \tag{77}$$

$$G(x_t, u_t) := \sum_{j=1}^{l} g_{j(1)}(x_t, u_t) g_{j(1)}(x_t, u_t)^{\mathrm{T}}. \tag{78}$$