2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

System Identification Based on Variational Bayes Method and The Invariance under Coordinate Transformations

Kenji Fujimoto, Akinori Satoh and Shuichi Fukunaga

Abstract— **This paper proposes a parameter estimation method for state-space models based on the variational Bayes method. We adopt the same form of functions as the prior and posterior probability distributions so that we can be used it iteratively to obtain accurate estimation whereas the existing algorithms cannot be used iteratively. Furthermore, the proposed algorithm is invariant under coordinate transformations, in the sense that the posterior probabilities of state-space models similar to each other are equivalent. Moreover, a numerical example demonstrates the effectiveness of the proposed method.**

I. Introduction

This paper is concerned with system identification based on the variational Bayes method [1]. System identification is a method to estimate system parameters of a given dynamical (state-space) system from the input and output data. Most popular system identification methods in control systems theory are based on the least square method which is related to the maximum likelihood estimation [2]. There is another famous approach called subspace system identification which is based on principal component analysis [3]. These methods provides deterministic estimation for state-space models.

On the other hand, stochastic estimation methods are often used in machine learning theory. The EM algorithm and the Bayesian inference are popular among them [4]. While both methods use a likelihood function as a system model, the former one gives deterministic estimations of the unknown parameters and the latter derives probability density functions of the unknown parameters. The Bayesian inference provides the reliability of the estimation and it also allows one to use the prior knowledge of the unknown parameters for the estimation. Such stochastic estimation is useful for control in order to suppress the variation of the transient behavior caused by the variation of the system parameters [5].

There exist some results on their application to state-space models. For instance, the paper [6] derived a system identification method based on the EM algorithm. Beal [7] proposes an identification algorithm based on the Bayesian inference by using its approximation called variational Bayes method [1]. The paper [8] also proposes a similar result based on the variational Bayes in which the state estimation algorithm is described by the Kalman filter and smoother. Although those results provide a new framework to estimate state-space models, they are incomplete because of the following two reasons: (a) Two estimated state-space systems which are transformed to each other by coordinate transformations have different probability distributions whereas their probability distributions should be the same because they have the same transfer functions. (b) The form of the prior distributions and that of the posterior ones are not the same so it is not possible to use them recursively.

The present paper proposes a novel system identification method based on the variational Bayes to overcome the problems mentioned above by adopting a more general class of state-space models than those treated in the existing results [7], [8]. Hence the proposed method is applicable to a wider class of systems compared to them. It is also noted that, in the proposed method, the form of the the posterior distributions are exactly same as that of the prior ones, i.e., they are conjugate priors [9]. This allows us to apply the proposed algorithm recursively to obtain a more accurate estimation than a single application of the algorithm. Hence the proposed method is both theoretically consistent and practically useful. Furthermore, numerical examples demonstrate the effectiveness of the proposed method. Most of the proofs in this paper are omitted due to the limitation of space. Please see [13] for the detail.

Let us define some notations used in this paper. A functional $\mathrm{KL}[p_1\|p_2] := \int p_1(x)\log(p_1(x)/p_2(x))\mathrm{d}x$ is defined for two arbitrary probability density functions $p_1(x)$ and $p_2(x)$ for a random variable $x \in \mathbb{R}^n$ which is called the Kullback-Leibler (KL) divergence. The expectation of a given function $f(x)$ subject to the probability density function of the random variable x is $p(x)$ is denoted by $\langle f(x)\rangle_{p(x)} := \int f(x)p(x)\mathrm{d}x$. The Gaussian distribution is denoted by $\mathcal{N}(x|\mu,\Sigma) := (2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\{-(1/2)(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\}$. The Wishart distribution is denoted by $\mathcal{W}(\Lambda|S,\nu) := C_{\mathcal{W}}\ |\Lambda|^{(\nu-n-1)/2}\exp\{-(1/2)\mathrm{tr}\{S^{-1}\Lambda\}\}$ where $C_{\mathcal{W}} = |S|^{-\nu/2}\{(2^{\nu n/2}\pi^{n(n-1)/4})\prod_{i=1}^{n}\Gamma((\nu+1-i)/2)\}^{-1}$.
The symbol vec denotes a function to produce a vector from a matrix by re-ordering its elements. For example, $\mathrm{vec}(B) = (b_1^{\mathrm{T}},\ldots,b_m^{\mathrm{T}})^{\mathrm{T}}$ holds for a matrix $B \in \mathbb{R}^{n\times m}$ whose i-th column is $b_i \in \mathbb{R}^n$.

II. Variational Bayes method

This section briefly reviews Bayesian inference [10] and the variational Bayes method [1].

K. Fujimoto is with Department of Mechanical Science and Engineering, Graduate School of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan fujimoto@nagoya-u.jp

A. Satoh is with Department of Mechanical Science and Engineering, Graduate School of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan a-satou@haya.nuem.nagoya-u.ac.jp

S. Fukunaga is with Tokyo Metropolitan College of Industrial Technology, 1-1-40 Higashioi, Shinagawa-ku, Tokyo, Japan fukunaga@s.metro-cit.ac.jp

978-1-61284-799-3/11/$26.00 ©2011 IEEE

A. Bayesian inference

The objective of Bayesian inference is to estimate the unknown parameter θ from the measured data Y. It is assumed that we know the conditional likelihood function $p(Y|\theta)$ of the measured data Y with respect to the condition θ is given. The estimation result is given by the the conditional probability function $p(\theta|Y)$ with respect to a given measured data Y which is called the *posterior distribution* of θ. Bayes' Theorem suggests that it is described by

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)}.$$

Here $p(\theta)$ is the probability density function of θ without using the measured data Y which is called the *prior distribution* of θ. $p(Y)$ is the marginal likelihood of Y satisfying

$$p(Y) = \int p(Y|\theta)p(\theta)\mathrm{d}\theta. \tag{1}$$

B. Variational Bayes method

In order to obtain the posterior distribution $p(\theta|Y)$, we need to calculate the integral in Equation (1) whose analytic solution is not known in general. There is an approach to compute it numerically by using the Monte Carlo method, e.g., [11]. But it requires a lot of computational effort. The variational Bayes method is a way to obtain an analytic solution to Equation (1) approximately. In what follows, we consider the case where there are two unknown parameters: the system parameter θ and the hidden variable X. The estimation of the true posterior distribution of those parameters $p(X, \theta|Y)$ is denoted by $q(X, \theta)$.

The marginal likelihood function $\log p(Y)$ is described by

$$\log p(Y) = \mathrm{F}[q(X,\theta)] + \mathrm{KL}[q(X,\theta)||p(X,\theta|Y)] \tag{2}$$

where $\mathrm{F}[\cdot]$ is a functional of $q(X, \theta)$ defined by

$$\mathrm{F}[q(X,\theta)] := \int q(X,\theta)\log\frac{p(Y,X|\theta)p(\theta)}{q(X,\theta)}\mathrm{d}X\mathrm{d}\theta.$$

Since the left hand side of Equation (2) does not depend on the estimated distribution q, maximizing $\mathrm{F}[q(X, \theta)]$ by selecting the variable q is equivalent to minimizing the KL divergence describing the distance between the true posterior distribution $p(X, \theta|Y)$ and its estimation $q(X, \theta)$. Consequently, a distribution $q(X, \theta)$ maximizing $\mathrm{F}[q(X, \theta)]$ is the best approximation of the true posterior distribution $p(X, \theta|Y)$.

Suppose that the system parameter is decomposed by independent components as $\theta = \{\theta_1, \ldots, \theta_I\}$ and that the true posterior distribution[1] is described by $p(X, \theta) = p(X)\prod_{i=1}^{I} p(\theta_i)$. Accordingly, its estimation can be written as $q(X, \theta) = q(X)\prod_{i=1}^{I} q(\theta_i)$. Then the functional $\mathrm{F}[q(X, \theta)]$ is described as

$$\mathrm{F}[q(X,\theta)] = \left\langle \log\frac{p(X,Y|\theta)}{q(X)}\right\rangle_{q(X),q(\theta)} + \sum_{i=1}^{I}\left\langle\log\frac{p(\theta_i)}{q(\theta_i)}\right\rangle_{q(\theta)}.$$

[1]For simplicity, the true posterior distribution $p(X, \theta|Y)$ is denoted by $p(X, \theta)$ in what follows.

Maximization of the functional F with the constraint $\int q(X,\theta)\mathrm{d}X\mathrm{d}\theta = 1$ is characterized by the extremum problem of the cost function $J[q(X)] = \mathrm{F}[q(X,\theta)] + \lambda(\int q(X,\theta)\mathrm{d}X\mathrm{d}\theta - 1)$ with a Lagrangian multiplier λ. A necessary condition of the solution q are characterized by the following Euler-Lagrange equations $(\delta J/\delta q) = 0$ and $(\delta J/\delta\lambda) = 0$. Solving them yields the following variational (estimated) posterior distribution of X.

$$q(X) = C_X \exp\langle\log p(X,Y|\theta)\rangle_{q(\theta)} \tag{3}$$

Here C_X is the normalizing constant to achieve $\int q(X)\mathrm{d}X = 1$. Similarly, we can obtain the variational (estimated) posterior distribution of θ_i as follows.

$$q(\theta_i) = C_{\theta_i} p(\theta_i)\exp\langle\log p(X,Y|\theta)\rangle_{q(X),q(\theta_{-i})} \tag{4}$$

Here C_{θ_i}'s are the normalizing constants and $\theta_{-i} = \{\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_I\}$.

Since the solutions $q(X)$ and $q(\theta_i)$'s satisfying Equations (3) and (4) cannot be solved analytically, an algorithm to compute them recursively as follows is adopted. Here k denotes the number of the iteration step and $q(\cdot)^{(k)}$ denotes the estimated posterior distribution at the step k.

Algorithm 1: (Variational Bayes method)
Step.1
Set the initial distribution $q(\theta)^{(0)}$ and $k = 0$.
Step.2
Compute the following steps:
VB-E step

$$q(X)^{(k+1)} = C_X \exp\langle\log p(X,Y|\theta)\rangle_{q(\theta)^{(k)}} \tag{5}$$

VB-M step

$$q(\theta_i)^{(k+1)} = C_{\theta_i} p(\theta_i) \times \exp\langle\log p(X,Y|\theta)\rangle_{q(X)^{(k+1)},q(\theta_{-i})^{(k)}} \tag{6}$$

Set $k \leftarrow k+1$ and repeat Step.2 until the solution converges.

III. Bayesian inference for state-space models

This section proposes a system identification method for state-space models based on the variational Bayes method.

A. Problem setting

Consider the following discrete-time linear system

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t + w_t \\ y_t &= Cx_t + Du_t + v_t \end{aligned}. \tag{7}$$

Here $x_t \in \mathbb{R}^n$ is the state, $u_t \in \mathbb{R}^m$ is the input, $y_t \in \mathbb{R}^l$ is the output, respectively. The matrices $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}, C \in \mathbb{R}^{l\times n}, D \in \mathbb{R}^{l\times m}$ are the system matrices. The distributions of $w_t \in \mathbb{R}^n$, $v_t \in \mathbb{R}^l$ and $x_0 \in \mathbb{R}^n$ are

$$p(w_t) = \mathcal{N}(w_t|0,Q), \quad p(v_t) = \mathcal{N}(v_t|0,R) \tag{8}$$

$$p(x_0) = \mathcal{N}(x_0|\mu_0, V). \tag{9}$$

The sequences of the state and the output are denoted by $X_N = \{x_0, \ldots, x_N\}$ and $Y_N = \{y_0, \ldots, y_N\}$, respectively. Then their log likelihood function is given as follows.

$$
\begin{aligned}
&\log\ p(X_N, Y_N) = \\
&-\frac{1}{2}\sum_{t=0}^{N-1}(x_{t+1}-Ax_t-Bu_t)^{\mathrm{T}}Q^{-1}(x_{t+1}-Ax_t-Bu_t) \\
&-\frac{1}{2}\sum_{t=0}^{N}(y_t-Cx_t-Du_t)^{\mathrm{T}}R^{-1}(y_t-Cx_t-Du_t) \\
&-\frac{1}{2}(x_0-\mu_0)^{\mathrm{T}}V^{-1}(x_0-\mu_0)-\frac{N}{2}\log|Q| \\
&-\frac{(n+l)(N+1)}{2}\log 2\pi-\frac{N+1}{2}\log|R|-\frac{1}{2}\log|V|
\end{aligned} \tag{10}
$$

The main objective of this paper is to obtain the estimation of the posterior distributions of the system parameters A, B, C, D, Q and R using their prior distributions.

B. Prior distributions

In order to execute the procedure of the variational Bayes method explained in Algorithm 1, we need to compute the integrals in Equations (5) and (6). There are a family of prior distributions called *conjugate priors* [9] for which there exist likelihood functions such that the posterior distributions have the same form as the prior ones. In the existing results on the variational Bayes methods for state-space models, the prior and posterior distributions do not coincide with each other exactly [7], [8]. This paper proposes the following set of the prior distributions and prove that the corresponding posterior distributions derived using the likelihood function (10) is the same form as the prior ones.

First of all, let us define the distributions of Q and R as

$$p(Q) = \mathcal{W}(Q^{-1}|S_Q, \nu) \tag{11}$$
$$p(R) = \mathcal{W}(R^{-1}|S_R, \eta). \tag{12}$$

The distributions of the parameters A, B, C and D are

$$p(A, B|Q) = \mathcal{N}(\mathrm{vec}(A, B)|\mu_{AB}, G \otimes Q) \tag{13}$$
$$p(C, D|R) = \mathcal{N}(\mathrm{vec}(C, D)|\mu_{CD}, H \otimes R) \tag{14}$$

where $\otimes$ denotes the Kronecker product.

C. Estimation of the system parameters

This subsection applies the VB-M step to the state-space model (7) and show how to compute the integral in (6). The estimated posterior distributions

$$
\begin{aligned}
&q(A, B, Q) \propto p(A, B, Q) \\
&\times \exp\langle \log p(X_N, Y_N)\rangle_{q(X_N)^{(k+1)}, q(\mathrm{vec}(C)\mathrm{vec}(D), R^{-1})^{(k)}}
\end{aligned} \tag{15}
$$

$$
\begin{aligned}
&q(C, D, R) \propto p(C, D, R) \\
&\times \exp\langle \log p(X_N, Y_N)\rangle_{q(X_N)^{(k+1)}, q(\mathrm{vec}(A), \mathrm{vec}(B), Q^{-1})^{(k)}}
\end{aligned} \tag{16}
$$

calculated using the prior distributions (11)–(14) and the likelihood function (10) are obtained as follows. Here $p(\cdot)$ and $q(\cdot)$ denote the prior and posterior distributions, respectively.

$$q(Q) = \mathcal{W}(Q^{-1}|\hat{S}_Q, \nu + N) \tag{17}$$
$$q(A, B|Q) = \mathcal{N}(\mathrm{vec}(A, B)|\hat{\mu}_{AB}, \hat{G} \otimes Q) \tag{18}$$

Similarly, we have

$$q(R) = \mathcal{W}(R^{-1}|\hat{S}_R, \eta + N) \tag{19}$$
$$q(C, D|R) = \mathcal{N}(\mathrm{vec}(C, D)|\hat{\mu}_{CD}, \hat{H} \otimes R). \tag{20}$$

The posterior distribution $q(A)$ is obtained by marginalizing Equation (18) with respect to B. $q(B)$, $q(C)$ and $q(D)$ are obtained in a similar way. Here $\hat{S}_{(\cdot)}, \hat{\mu}_{(\cdot)}, (\hat{\cdot}) \otimes (\hat{\cdot})$ are the estimation of the hyper parameter $S_{(\cdot)}$, the average $\mu_{(\cdot)}$ and the covariances $G \otimes Q$ and $H \otimes R$, respectively. The detail of the computation is given in Appendix.

D. Estimation of the states

This subsection applies VB-E Step to the state-space model (7) and calculate the integral (5). The state estimation procedure in the EM algorithm [12] giving the maximum likelihood estimation is same as the Kalman filter. Although VB-E Step of the variational Bayes method applied to a state-space model is different from the Kalman filter in general [7], it is proved that VB-E Step of a certain variational Bayes problem coincides with the Kalman filter and smoother of an augmented system of the original model (7) [8]. We adopt this framework and realize the algorithm by a Kalman filter and smoother. To this end, we start from computing the log likelihood function as follows.

$$
\begin{aligned}
&\langle \log p(X_N, Y_N)\rangle_{q(A,B,C,D,Q,R)^{(k)}} \\
&= -\frac{1}{2}\sum_{t=0}^{N-1}(x_{t+1}-\langle A\rangle x_t-\langle B\rangle u_t)^{\mathrm{T}}\langle Q^{-1}\rangle \\
&\quad\times(x_{t+1}-\langle A\rangle x_t-\langle B\rangle u_t)+\frac{N}{2}\langle\log|Q|\rangle_{q(Q)} \\
&\quad-\frac{1}{2}\sum_{t=0}^{N}(y_t-\langle C\rangle x_t-\langle D\rangle u_t)^{\mathrm{T}}\langle R^{-1}\rangle \\
&\quad(y_t-\langle C\rangle x_t-\langle D\rangle u_t)+\frac{N+1}{2}\langle\log|R|\rangle_{q(R^{-1})} \\
&\quad-\frac{1}{2}(x_0-\mu_0)^{\mathrm{T}}V^{-1}(x_0-\mu_0)\frac{1}{2}\log|V| \\
&\quad-\frac{(n+l)(N+1)}{2}\log 2\pi+\sum_{t=0}^{N}(x_t^{\mathrm{T}}\ u_t^{\mathrm{T}})S\begin{pmatrix}x_t\\u_t\end{pmatrix}
\end{aligned} \tag{21}
$$

Here the matrix S is defined using $\mathrm{cov}(A, B) := \langle A^{\mathrm{T}}QB\rangle - \langle A\rangle^{\mathrm{T}}\langle Q\rangle\langle B\rangle$ as follows.

$$
\begin{aligned}
S &= \begin{pmatrix}S_{11} & S_{12}\\ S_{21} & S_{22}\end{pmatrix} \\
S_{11} &= \mathrm{cov}(A, A) + \mathrm{cov}(C, C) \\
S_{12} &= S_{21}^{\mathrm{T}} = \mathrm{cov}(A, B) + \mathrm{cov}(C, D) \\
S_{22} &= \mathrm{cov}(B, B) + \mathrm{cov}(D, D)
\end{aligned}
$$

Suppose that $\mathrm{cov}(A,A) = 0$, $\mathrm{cov}(A,B) = 0$ and $\mathrm{cov}(B,B) = 0$ hold when $t = N$. Comparing Equations (21) and (10) we construct the augmented system of (21) as

$$\begin{aligned} x_{t+1} &= \tilde{A}x_t + \tilde{B}u_t + \tilde{w}_t, \qquad p(\tilde{w}) = \mathcal{N}(\tilde{w}|0,\tilde{Q}) \\ \tilde{y}_t &= \tilde{C}x_t + \tilde{D}u_t + \tilde{v}_t, \qquad p(\tilde{v}) = \mathcal{N}(\tilde{v}|0,\tilde{R}) \end{aligned} \tag{22}$$

where

$$\begin{aligned} &\tilde{A} = \langle A\rangle,\ \tilde{B} = \langle B\rangle,\ \tilde{Q} = \langle Q^{-1}\rangle^{-1} \\ &\tilde{y}_t = \begin{pmatrix} y_t \\ 0 \\ 0 \end{pmatrix},\ \tilde{C} = \begin{pmatrix} \langle C\rangle \\ L_{11} \\ L_{21} \end{pmatrix},\ \tilde{D} = \begin{pmatrix} \langle D\rangle \\ L_{12} \\ L_{22} \end{pmatrix} \\ &\tilde{R} = \mathrm{diag}(\langle R^{-1}\rangle^{-1}, I, I). \end{aligned}$$

The matrix L is obtained by the Cholesky decomposition of S as $S = L^{\mathrm{T}}L$. We construct the Kalman filter and smoother for the augmented system (22).

Let us design a Kalman filter for the system (22). The statistics of the initial state is defined by $\hat{x}_{0|0} = \mu_0, P_{0|0} = V$. The algorithm is summarized as follows where $t = 1,\ldots,N$. The estimate of the state x_t using the data with respect to the time $t = 1,\ldots,s$ is denoted by $\hat{x}_{t|s}$.

$$\begin{aligned} P_{t|t-1} &= \tilde{A}P_{t-1|t-1}\tilde{A}^{\mathrm{T}} + \tilde{Q} \qquad (23) \\ K_t &= P_{t|t-1}\tilde{C}^{\mathrm{T}}(\tilde{C}P_{t|t-1}\tilde{C}^{\mathrm{T}} + \tilde{R})^{-1} \\ P_{t|t} &= P_{t|t-1} - K_t\tilde{C}P_{t|t-1} \\ \hat{x}_{t|t-1} &= \tilde{A}\hat{x}_{t-1|t-1} + \tilde{B}u_{t-1} \\ \hat{x}_{t|t} &= \hat{x}_{t|t-1} + K_t(\tilde{y}_t - \tilde{C}\hat{x}_{t|t-1} - \tilde{D}u_t) \qquad (24) \end{aligned}$$

Next, the Kalman smoother is constructed as follows with $t = N-1,\ldots,0$ where the boundary conditions $\hat{x}_{N|N}$ and $P_{N|N}$ are same as the estimation of the Kalman filter.

$$\begin{aligned} S_t &= P_{t|t}\tilde{A}^{\mathrm{T}}P^{-1}_{t+1|t} \qquad (25) \\ \hat{x}_{t|N} &= \hat{x}_{t|t} + S_t[\hat{x}_{t+1|N} - \tilde{B}u_t - \tilde{A}\hat{x}_{t|t}] \\ P_{t|N} &= P_{t|t} + S_t[P_{t+1|N} - P_{t+1|t}]S_t^{\mathrm{T}} \qquad (26) \end{aligned}$$

The algorithm of VB-E Step is summarized as follows.

Algorithm 2: (Proposed method)

Step.1

Set the initial distributions $q(A,B|Q)^{(0)}, q(C,D|R)^{(0)}$ as in Equations (11)–(14) and $k = 0$.

Step.2

Compute the following steps:

VB-E Step

Calculate the estimation of the Kalman filter (23)–(24) and the Kalman smoother (25)–(26). Then the posterior distribution of x_t is obtained as follows.

$$q(x_t)^{(k+1)} = \mathcal{N}(x_t|\hat{x}_{t|N}, P_{t|N}), \quad t = 0,1,\ldots,N$$

VB-M Step

Calculate Equations (17)–(20) (with the equations in Appendix) to compute the following posterior distribution.

$$\begin{aligned} q(A,B|Q)^{(k+1)} &= \mathcal{N}(\mathrm{vec}(A,B)|\hat{\mu}_{AB}, \hat{G}\otimes Q) \qquad (27) \\ q(C,D|R)^{(k+1)} &= \mathcal{N}(\mathrm{vec}(C,D)|\hat{\mu}_{CD}, \hat{H}\otimes R) \qquad (28) \end{aligned}$$

Set $k \leftarrow k+1$ and repeat Step.2 until the solution converges.

Step.3

Marginalizing Equations (27) and (28), the posterior distributions of A, B, C and D are obtained.

IV. Invariance under coordinate transformations

The previous section proposes the algorithm to estimate the system parameters by the variational Bayes method. This section investigates the invariance of this algorithm under coordinate transformations and proves that the probabilities of two estimated state-space models which are transformed to each other by coordinate transformations are the same.

A. Problem setting

The system (7) is described on the new coordinate $\bar{x} = Tx$ as follows.

$$\begin{aligned} \bar{x}_{t+1} &= \bar{A}\bar{x}_t + \bar{B}u_t + \bar{w}_t, \quad p(\bar{w}_t) = \mathcal{N}(\bar{w}_t|0,\bar{Q}) \\ y_t &= \bar{C}\bar{x}_t + \bar{D}u_t + \bar{v}_t, \quad p(\bar{v}_t) = \mathcal{N}(\bar{v}_t|0,\bar{R}) \end{aligned} \tag{29}$$

The objective is to prove that the posterior distributions of the systems (7) and (29) are the same.

Let f_T denote the transformation of the system parameters corresponding to the coordinate transformation $\bar{x} = Tx$ described by

$$\begin{aligned} &(\bar{A},\bar{B},\bar{C},\bar{D},\bar{Q},\bar{R}) = f_T(A,B,C,D,Q,R) \\ &:= (f_T^A(A), f_T^B(B), f_T^C(C), f_T^D(D), f_T^Q(Q), f_T^R(R)) \\ &f_T^A(A) := TAT^{-1},\ f_T^B(B) := TB,\ f_T^C(C) := CT^{-1} \\ &f_T^D(D) := D,\ f_T^Q(Q) := TQT^{\mathrm{T}},\ f_T^R(R) := R. \end{aligned}$$

Let f_T^θ denote the transformation between the parameters $\theta = \{A,B,C,D\}$ and $\bar{\theta} = \{\bar{A},\bar{B},\bar{C},\bar{D}\}$ as

$$f_T^\theta(A,B,C,D) := (f_T^A(A), f_T^B(B), f_T^C(C), f_T^D(D)). \tag{30}$$

Let μ_A and Σ_{AA} denote the average and the covariance of the system parameters A. Then the average and the covariance of $\bar{A}$ are calculated as follows.

$$\begin{aligned} \mu_{\bar{A}} &= \mathrm{E}[\mathrm{vec}(\bar{A})] \\ &= \mathrm{E}[\mathrm{vec}(TAT^{-1})] \\ &= \mathrm{E}[(T^{-\mathrm{T}}\otimes T)\mathrm{vec}(A)] \\ &= (T^{-\mathrm{T}}\otimes T)\mathrm{E}[\mathrm{vec}(A)] = (T^{-\mathrm{T}}\otimes T)\mu_A \qquad (31) \\ \Sigma_{\bar{A}\bar{A}} &= \mathrm{E}[(\mathrm{vec}(\bar{A}) - \mathrm{E}[\mathrm{vec}(\bar{A})]) \\ &\quad \times (\mathrm{vec}(\bar{A}) - \mathrm{E}[\mathrm{vec}(\bar{A})])^{\mathrm{T}}] \\ &= \mathrm{E}[(T^{-\mathrm{T}}\otimes T)(\mathrm{vec}(A) - \mathrm{E}[\mathrm{vec}(A)]) \\ &\quad \times (\mathrm{vec}(A) - \mathrm{E}[\mathrm{vec}(A)])^{\mathrm{T}}(T^{-1}\otimes T^{\mathrm{T}})] \\ &= (T^{-\mathrm{T}}\otimes T)\mathrm{E}[(\mathrm{vec}(A) - \mathrm{E}[\mathrm{vec}(A)]) \\ &\quad \times (\mathrm{vec}(A) - \mathrm{E}[\mathrm{vec}(A)])^{\mathrm{T}}](T^{-1}\otimes T^{\mathrm{T}}) \\ &= (T^{-\mathrm{T}}\otimes T)\Sigma_{AA}(T^{-1}\otimes T^{\mathrm{T}}) \qquad (32) \end{aligned}$$

Note that Equation (13) implies that the statistics μ_A and Σ_{AA} have the relation to the parameters μ_{AB}, G and Q as follows.

$$\mu_{AB} := \begin{pmatrix} \mu_A^{\mathrm{T}} & \mu_B^{\mathrm{T}} \end{pmatrix}^{\mathrm{T}}, \quad G\otimes Q := \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}$$

The averages and covariances of the other parameters B, C and D are defined similarly to those of A, we can compute their statistics as follows.

$$\begin{aligned}
\mu_{\bar{B}} &= (I_u \otimes T)\mu_B \\
\Sigma_{\bar{B}\bar{B}} &= (I_u \otimes T)\Sigma_{BB}(I_u \otimes T^{\mathrm{T}}) \\
\mu_{\bar{C}} &= (T^{-\mathrm{T}} \otimes I_y)\mu_C \\
\Sigma_{\bar{C}\bar{C}} &= (T^{-\mathrm{T}} \otimes I_y)\Sigma_{CC}(T^{-1} \otimes I_y) \\
\mu_{\bar{D}} &= \mu_D \qquad (33) \\
\Sigma_{\bar{D}\bar{D}} &= \Sigma_{DD} \\
\Sigma_{\bar{A}\bar{B}} &= (T^{-\mathrm{T}} \otimes T)\Sigma_{AB}(I_u \otimes T^{\mathrm{T}}) \\
\Sigma_{\bar{C}\bar{D}} &= (T^{-\mathrm{T}} \otimes I_y)\Sigma_{CD} \qquad (34)
\end{aligned}$$

Here I_u and I_y are the identity matrices corresponding to the vectors u and y, respectively.

B. Invariance of the posterior distributions

This section proves the invariance of the posterior distributions under coordinate transformations. To this end, we start from the following lemma.

Lemma 1: The following relationship holds for the transformation (30) of the parameter θ.

$$\mathrm{d}\rho_{\bar{A}}\mathrm{d}\rho_{\bar{B}}\mathrm{d}\rho_{\bar{C}}\mathrm{d}\rho_{\bar{D}} = \mathrm{d}\rho_A\mathrm{d}\rho_B\mathrm{d}\rho_C\mathrm{d}\rho_D \tag{35}$$

Here ρ_A, ρ_B, ρ_C and ρ_D are the measures of the parameters A, B, C and D, respectively. They satisfy

$$\begin{aligned}
\mathrm{d}\rho_A &= \prod \mathrm{d}a_{i,j}, & \mathrm{d}\rho_B &= \prod \mathrm{d}b_{i,j} \\
\mathrm{d}\rho_C &= \prod \mathrm{d}c_{i,j}, & \mathrm{d}\rho_D &= \prod \mathrm{d}d_{i,j}
\end{aligned}$$

where $a_{i,j}$, $b_{i,j}$, $c_{i,j}$ and $d_{i,j}$ are the (i,j) elements of the matrices A, B, C and D, respectively.

Proof: Proof is omitted due to lack of space. ■

This lemma proves that the measures of the system parameters are invariant under coordinate transformations. The next theorem follows this lemma.

Theorem 1: The joint probability distribution of the system parameters A, B, C and D and that of the parameters $\bar{A}$, $\bar{B}$, $\bar{C}$ and $\bar{D}$ are equivalent, i.e.,

$$p(\bar{A}, \bar{B}, \bar{C}, \bar{D}) = p(A, B, C, D). \tag{36}$$

Proof: Proof is omitted for the reason of space. ■

This theorem shows that the probability distribution does not change for any coordinate transformation. Next, the relation between the Bayesian inference and the coordinate transformation is considered. To this end, the following assumption is employed.

Assumption 1: The prior distribution $p(A, B, C, D|Q, R)$ satisfies

$$p(A, B, C, D|Q, R) = \mathcal{N}(\mathrm{vec}(A, B, C, D)|\mu_{ABCD}, \Sigma) \tag{37}$$

where

$$\mu_{ABCD} := \begin{pmatrix} \mu_{AB}^{\mathrm{T}} & \mu_{CD}^{\mathrm{T}} \end{pmatrix}^{\mathrm{T}}, \Sigma := \begin{pmatrix} G \otimes Q & 0 \\ 0 & H \otimes R \end{pmatrix}. \tag{38}$$

In order to explain the result, we employ the following notations.

$$\begin{aligned}
T_{AB} &:= \begin{pmatrix} T^{-\mathrm{T}} \otimes T & 0 \\ 0 & I_u \otimes T \end{pmatrix} \\
T_{CD} &:= \begin{pmatrix} T^{-\mathrm{T}} \otimes I_y & 0 \\ 0 & I_u \otimes I_y \end{pmatrix} \\
T_{ABCD} &:= \begin{pmatrix} T_{AB} & 0 \\ 0 & T_{CD} \end{pmatrix}
\end{aligned}$$

Lemma 2: Consider the probability density function in Equation (37) in Assumption 1. The transformation (30) converts it to

$$p(\bar{A}, \bar{B}, \bar{C}, \bar{D}|\bar{Q}, \bar{R}) = \mathcal{N}(\mathrm{vec}(\bar{A}, \bar{B}, \bar{C}, \bar{D})|\mu_{\bar{A}\bar{B}\bar{C}\bar{D}}, \bar{\Sigma}) \tag{39}$$

where $\mu_{\bar{A}\bar{B}\bar{C}\bar{D}} := T_{ABCD}\ \mu_{ABCD}$ and $\bar{\Sigma} := T_{ABCD}\ \Sigma\ T_{ABCD}^{\mathrm{T}}$. Furthermore, we have

$$p(\bar{A}, \bar{B}, \bar{C}, \bar{D}|\bar{Q}, \bar{R}) = p(A, B, C, D|Q, R). \tag{40}$$

Proof: Proof is omitted for the limitation of the space. ■

Finally, the main result is stated as follows.

Theorem 2: Consider the variational Bayes algorithm in Algorithm 2. Let

$$\begin{aligned}
&q(A, B, C, D|Q, R) \\
&= \mathcal{N}(\mathrm{vec}(A, B, C, D)|\mu_{ABCD}, \Sigma)
\end{aligned}$$

denote the posterior distribution of θ derived using the initial prior distribution in Equation(37) in Assumption 1. Let

$$\begin{aligned}
&q(\bar{A}, \bar{B}, \bar{C}, \bar{D}|\bar{Q}, \bar{R}) \\
&= \mathcal{N}(\mathrm{vec}(\bar{A}, \bar{B}, \bar{C}, \bar{D})|\bar{\mu}_{ABCD}, \bar{\Sigma})
\end{aligned}$$

denote the posterior distribution of $\bar{\theta}$ derived using the initial prior distribution in Equation (39) in Lemma 2. Then we have

$$q(\bar{A}, \bar{B}, \bar{C}, \bar{D}|\bar{Q}, \bar{R}) = q(A, B, C, D|Q, R). \tag{41}$$

Furthermore, they satisfy

$$\begin{aligned}
&\int_{f_T^{\theta}(\Omega)} q(\bar{A}, \bar{B}, \bar{C}, \bar{D}|\bar{Q}, \bar{R})\mathrm{d}\rho_{\bar{A}}\mathrm{d}\rho_{\bar{B}}\mathrm{d}\rho_{\bar{C}}\mathrm{d}\rho_{\bar{D}} \\
&= \int_{\Omega} q(A, B, C, D|Q, R)\mathrm{d}\rho_A\mathrm{d}\rho_B\mathrm{d}\rho_C\mathrm{d}\rho_D \qquad (42)
\end{aligned}$$

for any region Ω.

Proof: Equation (41) is proved by substituting Equations (31)–(34) for Equation (42). Equation (42) follows from Lemmas 1 and 2. This completes the proof. ■

The theorem suggests that two estimated state-space realizations which can be transformed to each other have same probabilities. Namely, Algorithm 2 is invariant under the transformation f_T. This is a natural property of the state-space systems since input-output behavior of the state-space systems should be invariant under coordinate transformations. Furthermore, since the prior and posterior distributions have the same forms in the proposed algorithm, it can be used iteratively to obtain a more accurate estimation.

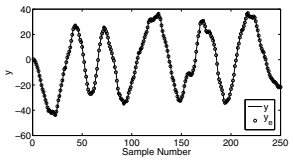

Fig. 1. Time responses of the output y and its estimation y_e

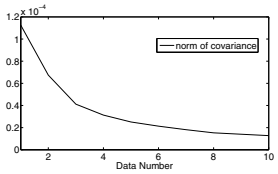

Fig. 2. History of the norm of the covariance matrix Σ

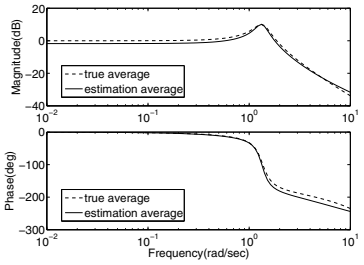

Fig. 3. Bode diagram of the true plant and its estimation

V. Numerical example

This sections gives a numerical example to exhibit how the proposed method in Algorithm 2 works.

A. Plant system

Consider the system (7) with the following parameters.

$$A = \begin{pmatrix} 1 & 0.3 \\ -0.06 & 0.94 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 \\ 0.06 \end{pmatrix} \tag{43}$$

$$C = \begin{pmatrix} 1 & 0 \end{pmatrix}, \qquad D = 0 \tag{44}$$

The terminal time is $N = 250$. We use 10 sets of input-output data. The averages of the noises w_t and v_t are zero and their covariance matrices are $Q = \mathrm{diag}(1, 1)$ and $R = 1$, respectively. The initial estimates are selected as follows.

$$\mu_0 = 0, \qquad V = 100I$$
$$\mu_{AB} = \begin{pmatrix} 0.1, & 0.1, & 0.1, & 0.1, & 0.1, & 0.1 \end{pmatrix}^{\mathrm{T}}$$
$$G = 100I, \qquad Q = 100I$$
$$\mu_{CD} = \begin{pmatrix} 0.1, & 0.1, & 0.1 \end{pmatrix}^{\mathrm{T}}$$
$$H = 100I, \qquad R = 100$$

The input signal is the M-sequence whose elements are within $[-50, 50]$.

B. Estimation result

Fig.1 shows the time responses of the measured (true) y and its estimate y_e computed by $y_{e,t} = \hat{\mu}_C \hat{x}_{t|N} + \hat{\mu}_D u_t$ where $\hat{\mu}_C$ and $\hat{\mu}_D$ are the averages of the posterior distributions of C and D, respectively. The solid line depicts the true y and the line with $\circ$ depicts its estimate y_e.

Fig.3 shows the bode diagrams of the true plant and its estimate. The dashed line depicts the bode diagram of the transfer function computed by the true system parameters A, B, C and D. The solid line depicts that computed by the averages of their estimations given as follows.

$$A = \begin{pmatrix} 0.7480 & -0.3735 \\ -0.1820 & 1.1994 \end{pmatrix}, \qquad B = \begin{pmatrix} 0.2847 \\ -0.0717 \end{pmatrix} \tag{45}$$

$$C = \begin{pmatrix} 0.0944 & 0.5175 \end{pmatrix}, \qquad D = -0.0070 \tag{46}$$

$$Q = \begin{pmatrix} 17.6189 & -9.3673 \\ -9.3673 & 9.9706 \end{pmatrix}, \qquad R = 1.0639$$

Fig.2 shows the history of the norm of the covariance matrix Σ of the parameter θ in Equation (38) along the number of the input-output data sets used for the estimation.

C. Comments

Fig.1 shows that the estimated output y_e is very close to the true output y. Fig.3 indicates that the transfer function of the estimated model is very similar to that of the true plant. These figures show that the proposed method gives a very accurate estimation of the plant model, although the parameters A, B, C and D and their estimations are not identical due to the freedom of the coordinate transformation. Fig.2 shows that the variance Σ decreases monotonically. This means that the reliability of the estimation increases as the number of the used data increases.

VI. Conclusion

This paper proposes a system identification method for linear discrete-time state-space systems based on the variational Bayes method. The invariance of the algorithm under coordinate transformations is proved, that is, two estimated systems which are converted to each other have the same probabilities in the proposed algorithm. Also, since the posterior distribution has the same form as the prior one, the proposed algorithm can be used iteratively to obtain an accurate estimation. Furthermore, a numerical example confirms the effectiveness of the proposed method.

References

[1] H. Attias, "Inferring parameters and structure of latent variable models by variational bayes," in *Proc. 15th Conf. on Uncertainty in Artificial Intelligence*, 1999, pp. 21–30.

[2] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, Upper Saddle River, New Jersey, second edition, 1999.

[3] T. Katayama, *Subspace Methods for System Identification*, Springer Verlag, London, 2005.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[5] K. Fujimoto, S. Ogawa, Y. Ota, and M. Nakayama, "Optimal control of linear systems with stochastic parameters for variance suppression: The finite time horizon case," in *Proceedings of the 18th IFAC World Congress*, 2011.

[6] S. Gibson and B. Ninness, "Robust maximum-likehood estimation of multivariable dynamic systems," *Automatica*, vol. 41, pp. 1667–1682, 2005.

[7] M. J. Beal, *Variational Algorithms for Approximate Bayesian inference*, Ph.D. thesis, University of Londong, London, UK, 2003.

[8] D. Barber and S. Chiappa, "Unified inference for variational Bayesian linear Gaussian state-space models," in *Advances in Neural Information Processing Systems 19 (NIPS 20)*, pp. 81–88. The MIT Press, 2007.

[9] D. Persi and Y. Donald, "Conjugate priors for exponential families," *The Annals of Statistics*, vol. 7, no. 2, pp. 269–281, 1979.

[10] G. E. P. Box, "Sampling and Bayes's inference in scientific modelling and robustness," *Journal of the Royal Statistical Society*, vol. 143, no. 4, pp. 383–430, 1980.

[11] B. Ninness and S. J. Henriksen, "Bayesian system identification via markov chain monte carlo techniques," *Automatica*, vol. 46, no. 1, pp. 40–51, 2010.
[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.
[13] K. Fujimoto, A. Satoh, and S. Fukunaga, "System identification based on variational bayes method and the inavriance under coordinate transformations," Submitted, 2011.

APPENDIX

Appendix shows the detailed calculation of the posterior distributions in Equations (17)–(20). First of all, marginalization of the distribution (18) is

$$q(\mathrm{vec}(A)|Q^{-1}) = \mathcal{N}(\mathrm{vec}(A)|\hat{\mu}_A, \hat{\Sigma}_A \otimes Q)$$
$$q(\mathrm{vec}(B)|Q^{-1}) = \mathcal{N}(\mathrm{vec}(B)|\hat{\mu}_B, \hat{\Sigma}_B \otimes Q).$$

Using the operator 'mat' denoting the inverse of 'vec,' we obtain

$$\begin{aligned}
\hat{S}_Q^{-1} &:= W_A + S_Q^{-1} + \mathrm{mat}(\mu_{AB})G^{-1}\{I_x \\
&\quad - (\Sigma_Q^{-1} + G^{-1})^{-1}G^{-1}\}\mathrm{mat}(\mu_{AB})^{\mathrm{T}} \\
&\quad - (S_A^{\mathrm{T}}\ \tilde{M}^{\mathrm{T}})G^{-1}\mathrm{mat}(\mu_{AB})^{\mathrm{T}} \\
&\quad - (S_A^{\mathrm{T}}\ \tilde{M}^{\mathrm{T}})(\Sigma_Q^{-1} + G^{-1})^{-1}(S_A^{\mathrm{T}}\ \tilde{M}^{\mathrm{T}})^{\mathrm{T}} \\
\hat{\Sigma}_A^{-1} &:= V_A + M_A - (G_A + M_C^{\mathrm{T}})(U_U + M_D)^{-1} \\
&\quad \times (G_A^{\mathrm{T}} + M_C) \qquad (47) \\
\hat{\mu}_A &:= (\Sigma_A \otimes I_x)([\{M_A - (G_A + M_C^{\mathrm{T}})(U_U + M_D)^{-1} \\
&\quad \times M_C\} \otimes I_x]\mu_A + [\{M_B - (G_A + M_C^{\mathrm{T}}) \\
&\quad \times (U_U + M_D)^{-1}M_D\} \otimes I_x]\mu_B + \mathrm{vec}(S_A^{\mathrm{T}}) \\
&\quad - \{(G_A + M_C^{\mathrm{T}})(U_U + M_D)^{-1} \otimes I_x\}\mathrm{vec}(\tilde{M}^{\mathrm{T}})) \\
&\qquad (48) \\
\hat{\Sigma}_B^{-1} &:= U_U + M_D - (G_A^{\mathrm{T}} + M_B^{\mathrm{T}})(V_A + M_A)^{-1} \\
&\quad \times (G_A + M_B) \\
\hat{\mu}_B &:= (\Sigma_B \otimes I_x)([\{M_C - (G_A^{\mathrm{T}} + M_B^{\mathrm{T}})(V_A + M_A)^{-1} \\
&\quad \times M_A\} \otimes I_x]\mu_A + [\{M_D - (G_A^{\mathrm{T}} + M_B^{\mathrm{T}}) \\
&\quad \times (V_A + M_A)^{-1}M_B\} \otimes I_x]\mu_B + \mathrm{vec}(\tilde{M}^{\mathrm{T}}) \\
&\quad + \{(G_A^{\mathrm{T}} + M_B^{\mathrm{T}})(U_U + M_D)^{-1} \otimes I_x\}\mathrm{vec}(S_A^{\mathrm{T}}))
\end{aligned}$$

$$\Sigma_Q^{-1} := \begin{bmatrix} V_A & G_A \\ G_A^{\mathrm{T}} & U_U \end{bmatrix} \qquad G^{-1} := \begin{bmatrix} M_A & M_B \\ M_C & M_D \end{bmatrix}.$$

Similarly, we have

$$q(R^{-1}) = \mathcal{W}(R^{-1}|\eta + N, \hat{S}_R)$$
$$q(\mathrm{vec}(C)|R^{-1}) = \mathcal{N}(\mathrm{vec}(C)|\hat{\mu}_C, \hat{\Sigma}_C \otimes R)$$
$$q(\mathrm{vec}(D)|R^{-1}) = \mathcal{N}(\mathrm{vec}D|\hat{\mu}_D, \hat{\Sigma}_D \otimes R)$$

$$\begin{aligned}
\hat{S}_R^{-1} &:= (Y_Y + S_R^{-1} - \mathrm{mat}(\mu_{CD})H^{-1}\{I \\
&\quad - (\Sigma_R^{-1} + H^{-1})^{-1}H^{-1}\}\mathrm{mat}(\mu_{CD})^{\mathrm{T}} \\
&\quad - (S_C^{\mathrm{T}}\ U_Y^{\mathrm{T}})H^{-1}\mathrm{mat}(\mu_{CD})^{\mathrm{T}} \\
&\quad - (S_C^{\mathrm{T}}\ U_Y^{\mathrm{T}})(\Sigma_R^{-1} + H^{-1})^{-1}(S_C^{\mathrm{T}}\ U_Y^{\mathrm{T}})^{\mathrm{T}}) \\
\hat{\Sigma}_C^{-1} &:= W_C + H_A - (G_C + H_C^{\mathrm{T}})(U_U' + H_D)^{-1} \\
&\quad \times (G_C^{\mathrm{T}} + H_C) \\
\hat{\mu}_C &:= (\Sigma_C \otimes I_y)([\{H_A - (G_C + H_C^{\mathrm{T}})(U_U' + H_D)^{-1} \\
&\quad \times H_C\} \otimes I_y]\mu_C + [\{H_B - (G_C + H_C^{\mathrm{T}}) \\
&\quad \times (U_U' + H_D)^{-1}H_D\} \otimes I_y]\mu_D + \mathrm{vec}(S_C^{\mathrm{T}}) \\
&\quad - \{(G_C + H_C^{\mathrm{T}})(U_U' + H_D)^{-1} \otimes I_y\}\mathrm{vec}(U_Y^{\mathrm{T}})) \\
\hat{\Sigma}_D^{-1} &:= U_U' + H_D - (G_C^{\mathrm{T}} + H_B^{\mathrm{T}})(W_C + H_A)^{-1} \\
&\quad \times (G_C + H_B) \\
\hat{\mu}_D &:= (\Sigma_D \otimes I_y)([\{H_C - (G_C^{\mathrm{T}} + H_B^{\mathrm{T}})(W_V + H_A)^{-1} \\
&\quad \times H_A\} \otimes I_y]\mu_C + [\{H_D - (G_C^{\mathrm{T}} + H_B^{\mathrm{T}}) \\
&\quad \times (W_C + H_A)^{-1}H_B\} \otimes I_y]\mu_D + \mathrm{vec}(U_Y^{\mathrm{T}}) \\
&\quad + \{(G_C^{\mathrm{T}} + H_B^{\mathrm{T}})(U_U' + H_D) \otimes I_y\}\mathrm{vec}(S_C^{\mathrm{T}}))
\end{aligned}$$

$$H^{-1} := \begin{bmatrix} H_A & H_B \\ H_C & H_D \end{bmatrix} \qquad \Sigma_R^{-1} := \begin{bmatrix} W_C & G_C \\ G_C^{\mathrm{T}} & U_U' \end{bmatrix}.$$

Here the symbols Y_Y, U_Y, U_U, U_U' denote the input and output data using Y_N, U_N as follows.

$$Y_Y := \sum_{t=0}^{N} y_t y_t^{\mathrm{T}}, \quad U_Y := \sum_{t=0}^{N} u_t y_t^{\mathrm{T}} \qquad (49)$$

$$U_U := \sum_{t=0}^{N-1} u_t u_t^{\mathrm{T}}, \quad U_U' := \sum_{t=0}^{N} u_t u_t^{\mathrm{T}} \qquad (50)$$

The symbols $V_A, W_A, G_A, \tilde{M}, S_A, W_C, G_C, S_C$ denote the sufficient statistics defined as follows.

$$V_A := \sum_{t=0}^{N-1} \langle x_t x_t^{\mathrm{T}}|Y\rangle = \sum_{t=0}^{N-1} \{\hat{x}_{t|N}\hat{x}_{t|N}^{\mathrm{T}} + P_{t|N}\} \qquad (51)$$

$$W_A := \sum_{t=0}^{N-1} \langle x_{t+1} x_{t+1}^{\mathrm{T}}|Y\rangle = \sum_{t=0}^{N-1} \{\hat{x}_{t+1|N}\hat{x}_{t+1|N}^{\mathrm{T}} + P_{t+1|N}\}$$

$$G_A := \sum_{t=0}^{N-1} \langle x_t|Y\rangle u_t^{\mathrm{T}} = \sum_{t=0}^{N-1} \hat{x}_{t|N} u_t^{\mathrm{T}}$$

$$\tilde{M} := \sum_{t=0}^{N-1} u_t \langle x_{t+1}|Y\rangle^{\mathrm{T}} = \sum_{t=0}^{N-1} u_t \hat{x}_{t+1|N}^{\mathrm{T}}$$

$$S_A := \sum_{t=0}^{N-1} \langle x_t x_{t+1}^{\mathrm{T}}|Y\rangle = \sum_{t=0}^{N-1} \{\hat{x_{t|N}}\hat{x}_{t+1|N}^{\mathrm{T}} + M_{t+1|t}\}$$

$$W_C := \sum_{t=0}^{N} \langle x_t x_t^{\mathrm{T}}|Y\rangle = \sum_{t=0}^{N} \{\hat{x}_{t|N}\hat{x}_{t|N}^{\mathrm{T}} + P_{t|N}\}$$

$$G_C := \sum_{t=0}^{N} \langle x_t|Y\rangle u_t^{\mathrm{T}} = \sum_{t=0}^{N} \hat{x}_{t|N} u_t^{\mathrm{T}}$$

$$S_C := \sum_{t=0}^{N} \langle x_t|Y\rangle y_t^{\mathrm{T}} = \sum_{t=0}^{N} \hat{x}_{t|N} y_t^{\mathrm{T}} \qquad (52)$$

Here $M_{t+1|t}$ is the outcome of the Kalman smoother with the initial condition $M_{N|N-1} = (I - K_N\tilde{C})\tilde{A}P_{N-1|N-1}$ and compute it from $t = N$ to $t = 1$ as follows.

$$M_{t|t-1} = P_{t|t}S_{t-1}^{\mathrm{T}} + S_t(M_{t+1|t} - \tilde{A}P_{t|t})S_{t-1}^{\mathrm{T}}$$