IEICE TRANS. INF. & SYST., VOL.E97–D, NO.4 APRIL 2014

PAPER

Asymptotic Marginal Likelihood on Linear Dynamical Systems

Takuto NAITO[†], *Nonmember and* **Keisuke YAMAZAKI**[††a)], *Member*

SUMMARY Linear dynamical systems are basic state space models literally dealing with underlying system dynamics on the basis of linear state space equations. When the model is employed for time-series data analysis, the system identification, which detects the dimension of hidden state variables, is one of the most important tasks. Recently, it has been found that the model has singularities in the parameter space, which implies that analysis for adverse effects of the singularities is necessary for precise identification. However, the singularities in the models have not been thoroughly studied. There is a previous work, which dealt with the simplest case; the hidden state and the observation variables are both one dimensional. The present paper extends the setting to general dimensions and more rigorously reveals the structure of singularities. The results provide the asymptotic forms of the generalization error and the marginal likelihood, which are often used as criteria for the system identification.

key words: *Bayesian learning, Kalman filter, time-series data analysis*

1. Introduction

Linear dynamical systems (LDS) are basic state space models employed for modeling practical complex systems, time-series analysis, and image processing [1]–[3]. For practical usage, system identification is one of the most important modeling tasks. If the system is not identified well, which means that a dimension of the hidden state variables is not properly determined, estimated dynamics of inner states are not informative. There are active and passive approaches to the identification. For example, the frequency response analysis, which controls input signals and analyzes the output responses, is a representative active approach. When the LDS are used for time-series analysis, we often cannot affect the system, i.e. control of the input signal is not straightforward. Then, the system must passively be detected from the given observable data. This procedure corresponds to model selection in statistics.

From the statistical point of view, the LDS are one of parametric models, the parameters of which are expressed as coefficients of state space equations. The Kalman filter [4] is the most popular algorithm to derive values of hidden state variables from the observations when the coefficients of the equations are all given. The present paper focuses on the cases, where the parameters are unknown and the parameter learning is necessary. The parametric models fall into two kinds: regular and singular. The model is referred to as regular if there is one-to-one relation between the parameters and the model expression as a probabilistic function. Otherwise, it is singular. Properties of the parameter learning depend on whether the model is regular or not. Moreover, the conventional statistical methods for the model selection are not theoretically applicable to singular models.

Recently, singular models have been studied in the Bayes statistics. A relation between performance of the Bayesian learning and algebraic geometry has been found [5], [6]. The singularities in the parameter space play an important role to determine the performance. Mathematical approaches to reveal the structures of singularities have been developed [7], [8]. Based on the algebraic geometrical method, singularities in many models have been analyzed [9]–[17]. The structure of singularities depends on a model, which means that each model requires its own analysis. For example, a mathematical technique is an eigen value analysis in the reduced rank approximation [16] while it is an ideal-theoretic approach in the general Markov model [17].

A previous study [18] pointed out that LDS are singular when there is redundancy on the hidden state variables. It dealt with the simplest model; both the hidden state and the observation variables are one dimensional. The result shows that the Kalman filter can derive adverse hidden state estimation due to singularities caused by the redundancy of the hidden state variable. Therefore, more precise analysis of singularities on general setting of the model is necessary for appropriate system identification.

The present paper extends the model setting to general dimensions of the variables, and clarifies detailed structure of singularities. More precisely, the asymptotic forms of the marginal likelihood and the generalization error, which are representative criteria for the model selection, are derived. In regular models, the criteria AIC [19] and BIC [20] are based on the asymptotic forms of the generalization error, and the marginal likelihood, respectively. In singular models, on the other hand, they do not have theoretical validity. To tackle this issue, some criteria have been proposed [21]–[23]. The results of the present paper also provide helpful insights for managing these criteria.

The remainder of the paper is organized as follows. Section 2 introduces the Bayesian learning connected to the algebraic geometrical method. Section 3 describes the structure of singularities as theorems, proofs of which are in

Manuscript received June 17, 2013.
Manuscript revised November 25, 2013.

[†]The author is with the AVC Networks Company, Panasonic Corporation, Kadoma-shi, 571–8504 Japan.
[††]The author is with Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, Yokohama-shi, 226–8503 Japan.
a) E-mail: k-yam@math.dis.titech.ac.jp

DOI: 10.1587/transinf.E97.D.884

Sect. 4. Sections 5 and 6 state discussions and our conclusions, respectively.

2. Bayesian Parameter Learning and Analysis of Singularities

This section introduces the Bayesian learning and the connection to algebraic geometry.

Let $X^n = (X_1, X_2, \ldots, X_n)$ be a set of training samples taken independently and identically from the true distribution $q(X)$, where n is the number of training samples. Because LDS deal with sequential data, each X_i $(i = 1, \ldots, n)$ is a sequence whose length is T, i.e. $X_i = (x_1^i, \ldots, x_t^i, \ldots, x_T^i)$. Note that the sequence data X^n are taken as i.i.d. whereas each sequence X_i is not. Let $p(X|\omega)$ be a learning model, and $\varphi(\omega)$ be a prior distribution. The posterior distribution is given by

$$p(\omega|X^n) = \frac{1}{Z(X^n)} L(\omega)\varphi(\omega), \tag{1}$$

where the likelihood is defined by

$$L(\omega) = \prod_{i=1}^{n} p(X_i|\omega) \tag{2}$$

and the normalizing constant $Z(X^n)$ is the marginal likelihood. In practical situations, the minus log marginal likelihood $-\ln Z(X^n)$ is widely used for the model selection. Let us define the Bayesian free energy $F(n)$,

$$F(n) = E_{X^n}\Big[-\ln Z(X^n) + \sum_{i=1}^{n} \ln q(X_i)\Big], \tag{3}$$

where $E_{X^n}[\cdot]$ stands for the expectation over X^n. Note that $F(n)$ averagely reflects behavior of $-\ln Z(X^n)$ because the second term in the expectation is independent of the learning model and the prior.

In the Bayes method, the predictive distribution for unknown data X is constructed as

$$p(X|X^n) = \int p(X|\omega)p(\omega|X^n)d\omega. \tag{4}$$

In the maximum likelihood and the maximum a posteriori (MAP) methods, it is given by

$$p(X|X^n) = p(X|\omega_{ML}), \tag{5}$$
$$p(X|X^n) = p(X|\omega_{MAP}), \tag{6}$$

respectively, where the estimators are defined by

$$\omega_{ML} = \arg\max_{\omega} L(\omega), \tag{7}$$
$$\omega_{MAP} = \arg\max_{\omega} L(\omega)\varphi(\omega). \tag{8}$$

The performance of the model is measured by the difference of the distributions,

$$G(n) = E_{X^n}\Big[\int q(X) \ln \frac{q(X)}{p(X|X^n)} dX\Big], \tag{9}$$

which is the generalization error. In the Bayes method, the generalization error has the following relation to $F(n)$,

$$G(n) = F(n+1) - F(n). \tag{10}$$

This implies that the asymptotic form of $G(n)$ is naturally derived from that of $F(n)$.

Combination of the Mellin and the Laplace transforms changes $F(n)$ into the zeta function given by

$$\zeta(z) = \int H(\omega)^z \varphi(\omega) d\omega, \tag{11}$$
$$H(\omega) = \int q(X) \ln \frac{q(X)}{p(X|\omega)} dX. \tag{12}$$

In algebraic analysis, it is ensured that the zeta function has only real negative and rational poles. Let $0 > -\lambda_1 > -\lambda_2 > \ldots$ be a sequence of the poles, and $m_1, m_2, \ldots$ be the respective orders. The inverse transforms derive the asymptotic form of the energy function,

$$F(n) = \lambda_1 \ln n - (m_1 - 1) \ln \ln n + O(1) \tag{13}$$

for $n \to \infty$. Based on the relation Eq. (10), the generalization error has the asymptotic form,

$$G(n) = \frac{\lambda_1}{n} - \frac{m_1 - 1}{n \ln n} + o\Big(\frac{1}{n \ln n}\Big). \tag{14}$$

In the regular models, it has been proved that $\lambda_1 = \dim \omega/2$ and $m_1 = 1$, i.e.,

$$F(n) = \frac{\dim \omega}{2} \ln n + O(1), \tag{15}$$
$$G(n) = \frac{\dim \omega}{2n} + o\Big(\frac{1}{n \ln n}\Big). \tag{16}$$

Note that the coefficients only depend on the dimension of parameters. The well-known criteria BIC and AIC are derived on the basis of these asymptotic forms to select the optimal model [19], [20].

In the singular models, as seen in Eqs. (13) and (14), analyzing singularities via the zeta function plays an important role to know asymptotic behavior of the main objective functions for the model selection. Therefore, the present paper provides precise calculation of λ_1 and m_1 in LDS.

In many cases, finding the largest pole requires complicated mathematical calculation such as [17], [24]. When a pole $z = -\lambda$ and its order m have been calculated, upper bounds are derived as

$$F(n) \le \lambda \ln n - (m-1) \ln \ln n + O(1), \tag{17}$$
$$G(n) \le \frac{\lambda}{n} - \frac{m-1}{n \ln n} + o\Big(\frac{1}{n \ln n}\Big). \tag{18}$$

3. Analysis for Structure of the Singularities

This section shows analysis results of singularities in the parameter space. First, the learning and the true models are formulated as state space equations. Next, poles of the zeta

function, which determine the asymptotic forms of the generalization error and the marginal likelihood, are described as the main theorems.

3.1 Formulation of Linear Dynamical Systems

Let $z_t \in \mathbf{R}^q$ and $x_t \in \mathbf{R}^p$ be the hidden state and the output vectors at time t, respectively. The process and the observation noises are given by $w_t \in \mathbf{R}^q$ and $v_t \in \mathbf{R}^p$, respectively. LDS have the following state space equations,

$$z_{t+1} = Az_t + Dw_t, \tag{19}$$

$$x_t = Cz_t + v_t, \tag{20}$$

where $A \in \mathbf{R}^{q\times q}$ is a state matrix, $C \in \mathbf{R}^{p\times q}$ is an output matrix, and the elements of $D \in \mathbf{R}^{q\times q}$ are the coefficients of the process noise. The noises are assumed to follow a standard normal distribution.

Computing the likelihood $L(\omega)$ is the most necessary in any parameter learning. As seen in the predictive distribution and the posterior distribution, the Bayes learning also requires a value of the likelihood for the given ω. The Kalman filter effectively constructs the computing algorithm in LDS. The filter consists of two steps:

$$\hat{z}_{t|t-1} = A\hat{z}_{t-1|t-1}, \tag{21}$$

$$P_{t|t-1} = AP_{t-1|t-1}A^\top + DD^\top \tag{22}$$

for the predicting step, and

$$K_t = P_{t|t-1}C^\top \left(I + CP_{t|t-1}C^\top\right)^{-1}, \tag{23}$$

$$\hat{z}_{t|t} = \hat{z}_{t|t-1} + K_t\left(x_t - C\hat{z}_{t|t-1}\right), \tag{24}$$

$$P_{t|t} = \left(I - K_tC\right)P_{t|t-1} \tag{25}$$

for the updating step. The matrix I is a unit matrix and K_t is called the Kalman gain.

The model probability $p(X|\omega)$, where the parameters $\omega = (A, C, D, z_1)$, is expressed as

$$p(X|\omega) = p(x_1|\omega)\prod_{t=2}^{T} p(x_t|x_1, \ldots, x_{t-1}, \omega). \tag{26}$$

Based on properties of the normal distribution, the model probability can be rewritten as

$$p(X|\omega) = \prod_{t=1}^{T} \mathcal{N}(x_t|C\hat{z}_{t|t-1}, I + CP_{t|t-1}C^\top). \tag{27}$$

where the initial state setting is defined as $\hat{z}_{1|0} = z_1$ and $P_{1|0} = 0$. The likelihood is calculated as

$$L(\omega) = \prod_{i=1}^{n} p(X_i|\omega) = \prod_{i=1}^{n}\prod_{t=1}^{T} \mathcal{N}(x_t^i|C\hat{z}_{t|t-1}^i, I + CP_{t|t-1}^iC^\top) \tag{28}$$

where $\hat{z}_{t|t-1}^i$ and $P_{t|t-1}^i$ are iteratively evaluated on the basis of the Kalman filter.

Assume that the true model is defined as

$$q(X) = \prod_{t=1}^{T} \mathcal{N}(x_t|0, I), \tag{29}$$

where 0 and I are p dimensional zero vector and the unit matrix, respectively. The true model corresponding to no-hidden-state model generates i.i.d. data while the learning model treats them as the time-dependent data. This setting extracts a basic structure of singularities in more general cases, where the true model has q^*-dimensional hidden state vector, i.e. $z_t \in \mathbf{R}^{q^*}$. Availability of the setting will again be discussed in Sect. 5.

The length of sequences T affects the learning result. In the present paper, we assume a sufficient long sequences for complete parameter learning such as $T > \dim \omega$.

3.2 The Main Results

The following results show poles of the zeta function. The proofs will appear in the next section.

Theorem 1: When the true model and a learning model are defined by Eq. (29) and Eqs. (19)–(20), respectively, the zeta function has poles:

$$\lambda = \frac{q}{2}\min\left\{p, \frac{3}{2}q + 1\right\}, \tag{30}$$

$$m = \begin{cases} 2 & (p = \frac{3}{2}q + 1), \\ 1 & (\text{others}) \end{cases}. \tag{31}$$

The following larger poles are obtained when the initial state is at the origin;

Theorem 2: When the true model and a learning model are defined by Eq. (29) and Eqs. (19)–(20), respectively, the zeta function has poles:

$$\lambda = \frac{q}{4}\min\{p, q\}, \tag{32}$$

$$m = \begin{cases} 2 & (p = q), \\ 1 & (\text{others}) \end{cases}, \tag{33}$$

where the initial state is given as $z_1 = 0$.

Under the same conditions as [18], we obtain the exact expression instead of the bounds;

Corollary 1: When the true model and a learning model are defined by Eq. (29) and Eqs. (19)–(20), respectively, the zeta function has the largest pole:

$$\lambda_1 = \frac{1}{4}, \tag{34}$$

$$m_1 = 2, \tag{35}$$

where $z_1 = 0$, and $p = q = 1$.

4. Proofs of the Results

This section shows the proofs of the main results. First,

Sect. 4.1 shows some basic lemmas employed for the proofs. Then, the following Sects. 4.2, 4.3 and 4.4 prove Theorems 1 and 2 and Corollary 1, respectively.

4.1 Basic Lemmas

The following lemma is used for the proofs;

Lemma 1: For a positive constant $\epsilon < 1$, let $d \times d$ matrix satisfy $\|\Delta\| < \epsilon$, where $\|\cdot\|$ is an arbitrary norm. It holds that

$$\mathrm{Tr}(I+\Delta)^{-1} + \ln\det(I+\Delta) = d + \frac{1}{2}\mathrm{Tr}\Delta^2 + \mathrm{Tr}g(\Delta), \tag{36}$$

where $g(M)$ is a matrix polynomial function consisting of higher order terms than M^2

Proof: Based on $\ln\det M = \mathrm{Tr}\ln M$,

$$\mathrm{Tr}(I+\Delta)^{-1} + \ln\det(I+\Delta) = \mathrm{Tr}((I+\Delta)^{-1} + \ln(I+\Delta)). \tag{37}$$

By using the Taylor expansion with respect to Δ,

$$\begin{aligned}&\mathrm{Tr}(I+\Delta)^{-1} + \ln\det(I+\Delta)\\ &= \mathrm{Tr}(I - \Delta + \Delta^2 - \Delta^3 + \cdots + \Delta - \frac{1}{2}\Delta^2 + \frac{1}{3}\Delta^3 + \cdots)\\ &= d + \frac{1}{2}\mathrm{Tr}\Delta^2 + \mathrm{Tr}g(\Delta),\end{aligned} \tag{38}$$

which completes the proof. **(End of Proof)**

Let us introduce useful lemmas in the algebraic geometrical method without the proofs (cf. [25] for mathematical details).

Lemma 2: Let $\zeta_W(z)$ be a zeta function with a restriction of the parameter area;

$$\zeta_W(z) = \int_W H(\omega)^z \varphi(\omega) d\omega. \tag{39}$$

A pole $z = -\lambda_W$ and its order m_W of $\zeta_W(z)$ provide the following upper bounds,

$$F(n) \leq \lambda_W \ln n - m_W \ln\ln n + O(1). \tag{40}$$

Lemma 3: Let $H_u(\omega)$ be a function such that $H(\omega) \leq H_u(\omega)$ on the support of $\varphi(\omega)$. A pole $z = -\lambda_u$ and its order m_u in a zeta function $\zeta_u(z) = \int H_u(\omega)\varphi(\omega)d\omega$ provide the following upper bounds,

$$F(n) \leq \lambda_u \ln n - m_u \ln\ln n + O(1). \tag{41}$$

4.2 Proof of Theorem 1

Because it is not straightforward to calculate the integral of $H(\omega)$ on X, we will first find a polynomial of ω bounding $H(\omega)$ in two restricted areas of the parameter space. Based on Lemma 2–3, a pole of the zeta function with respect to the poly nominal is a bound of λ_1. Next, applying the blow-up to $H(\omega)$, we can change the polynomial into a monomial form of the parameter, which makes calculation of the integral in the zeta function easier. After marginalizing out the parameter, we will obtain a rational form of z and then find a pole in each restricted parameter area. Comparing the poles, we select the maximum pole as the tighter bound.

According to Eqs. (22), (23) and (25), there is a matrix $Q_t^{(1)}$ consisting of ω such that

$$P_{t|t-1} = AQ_t^{(1)}A^\top + DD^\top, \tag{42}$$

$$Q_t^{(1)} = (I - P_{t-1|t-2}C^\top(I + CP_{t-1|t-2}C^\top)^{-1}C)P_{t-1|t-2}. \tag{43}$$

By using $P_{t-1|t-2}$ and Eqs. (21), (23) and (24),

$$\begin{aligned}\hat{z}_{t|t-1} =& A(R_t^{(1)}\hat{z}_{t-1|t-2}\\ &+ P_{t-1|t-2}C^\top(I + CP_{t-1|t-2}C^\top)^{-1}x_{t-1})\end{aligned} \tag{44}$$

$$= Ay_t^{(1)}, \tag{45}$$

$$R_t^{(1)} = I - P_{t-1|t-2}C^\top(I + CP_{t-1|t-2}C^\top)^{-1}C, \tag{46}$$

where $y_t^{(1)}$ is a q dimensional vector consisting of ω and $x_1, \ldots, x_{t-1}$. Because the initial state is given as z_1,

$$P_{1|0} = 0, \tag{47}$$

$$\hat{z}_{1|0} = z_1. \tag{48}$$

Due to the expression Eq. (27),

$$p(X|\omega) = \mathcal{N}(x_1|Cz_1, I)\prod_{t=2}^{T}\mathcal{N}(x_t|CAy_t^{(1)}, \Sigma_t^{(1)}), \tag{49}$$

$$\Sigma_t^{(1)} = I + C(AQ_t^{(1)}A^\top + DD^\top)C^\top \tag{50}$$

Then, the Kullback information Eq. (12) is written as

$$\begin{aligned}H(\omega) =& E\Big[\ln\frac{\mathcal{N}(x_1|0,I)}{\mathcal{N}(x_1|Cz_1,I)}\Big]\\ &+ \sum_{t=2}^{T}E\Big[\ln\frac{\mathcal{N}(x_t|0,I)}{\mathcal{N}(x_t|CAy_t^{(1)},\Sigma_t^{(1)})}\Big],\end{aligned} \tag{51}$$

where $E[\cdot]$ denotes $\int \cdot q(X)dX$. On the covariance matrix, $E[y_t^{(1)}x_t^\top] = 0$ because $y_t^{(1)}$ consists of $x_1, \ldots, x_{t-1}$, which are independent of x_t due to the definition of $q(X)$. Then,

$$\begin{aligned}H(\omega) =& \frac{1}{2}z_1^\top C^\top Cz_1\\ &+ \sum_{t=2}^{T}\Big[\frac{1}{2}\ln\det\Sigma_t^{(1)} - \frac{p}{2} + \frac{1}{2}\mathrm{Tr}\Sigma_t^{(1)-1}\\ &+ \frac{1}{2}\mathrm{Tr}(CA)^\top\Sigma_t^{(1)-1}CAS_t^{(1)}\Big],\end{aligned} \tag{52}$$

where $S_t^{(1)} = E[y_t^{(1)}y_t^{(1)\top}]$. The elements of the parameter are denoted by

$$A = \{a_{ij}\}, \tag{53}$$

$$C = \{c_{ij}\}, \tag{54}$$
$$D = \{d_{ij}\}, \tag{55}$$
$$z_1 = (z_{11}, \ldots, z_{1q})^\top. \tag{56}$$

Let us restrict the parameter areas to $W_1^{(1)}$ and $W_2^{(1)}$. such that

$$W_1^{(1)} = \{\omega : \|c_{ij}\| < \epsilon^{(1)}\} \tag{57}$$
$$W_2^{(1)} = \{\omega : \|a_{ij}\| < \epsilon^{(1)}, \|d_{ij}\| < \epsilon^{(1)} \text{and} \|z_{1i}\| < \epsilon^{(1)}\}, \tag{58}$$

where $\epsilon^{(1)}$ is a small positive constant. Substituting Δ for $C(AQ_t^{(1)}A^\top + DD^\top)C^\top$ in Lemma 1, we derive that there is a positive constant $\alpha_1^{(1)}$ such that

$$H(\omega) \leq \alpha_1^{(1)} H_u^{(1)}(\omega) \tag{59}$$
$$\begin{aligned} H_u^{(1)}(\omega) =& \frac{1}{2} z_1^\top C^\top C z_1 \\ &+ \sum_{t=2}^{T} \Big[\frac{1}{4} (C(AQ_t^{(1)}A^\top + DD^\top)C^\top)^2 \\ &+ \frac{1}{2} \mathrm{Tr}(CA)^\top \Sigma_t^{(1)-1} CAS_t^{(1)} \Big] \end{aligned} \tag{60}$$

in the both area $W_1^{(1)}$ and $W_2^{(1)}$. The function $H_u^{(1)}(\omega)$ consists of terms $c_{ij}c_{kl}$ or terms $a_{ij}a_{kl}$, $d_{i_1i_2}d_{i_3i_4}d_{i_5i_6}d_{i_7i_8}$ and $z_{1i}z_{1j}$. In the restricted areas, it holds that

$$H(\omega) \leq \sum_{i,k=1}^{p} \sum_{j,l=1}^{q} c_{ij}c_{kl} f_{1ijkl}^{(1)}(\omega), \tag{61}$$
$$\begin{aligned} H(\omega) \leq & \Big\{ \sum_{i,k=1}^{q} \sum_{j,l=1}^{q} a_{ij}a_{kl} f_{2ijkl}^{(1)}(\omega) \\ &+ \sum_{i_1,\ldots,i_8=1}^{q} d_{i_1i_2}d_{i_3i_4}d_{i_5i_6}d_{i_7i_8} f_{3i_1\ldots i_8}^{(1)}(\omega) \\ &+ \sum_{i,j}^{q} z_{1i}z_{1j} f_{4ij}^{(1)}(\omega) \Big\}, \end{aligned} \tag{62}$$

respectively, where $f_{1ijkl}^{(1)}(\omega)$, $f_{2ijkl}^{(1)}(\omega)$, $f_{3i_1\ldots i_8}^{(1)}(\omega)$ and $f_{4ij}^{(1)}(\omega)$ are polynomials of ω.

For Eq. (61), we can find the following blow-up $\omega = \Phi_1^{(1)}(\hat{\omega})$;

$$c_{11} = \hat{c}_{11}, \tag{63}$$
$$c_{ij} = \hat{c}_{11}\hat{c}_{ij} \quad \text{(others)}. \tag{64}$$

Based on Lemmas 2 and 3, a pole of the zeta function $\zeta_1^{(1)}(z)$ provides upper bounds of the free energy;

$$\zeta_1^{(1)}(z) = \int_{W_1^{(1)}} \Big\{ \hat{c}_{11}^2 f_5^{(1)}(\hat{\omega}) \Big\}^z \varphi(\Phi_1^{(1)}(\hat{\omega}))|\Phi_1^{(1)}| d\hat{\omega}, \tag{65}$$

where $f_5^{(1)}(\hat{\omega})$ is a polynomial of $\hat{\omega}$ and $|\Phi_1^{(1)}|$ stands for the Jacobian. Because $|\Phi_1^{(1)}| = \hat{c}_{11}^{pq-1}$, the zeta function $\zeta_1^{(1)}$ has a pole at $z = -pq/2$.

For Eq. (62), we can find the following blow-up $\omega = \Phi_2^{(1)}(\hat{\omega})$;

$$a_{ij} = \hat{d}_{11}^2 \hat{a}_{ij} \quad (1 \leq i, j \leq q), \tag{66}$$
$$d_{11} = \hat{d}_{11}, \tag{67}$$
$$d_{ij} = \hat{d}_{11}\hat{d}_{ij} \quad \text{(others)}, \tag{68}$$
$$z_{1i} = \hat{d}_{11}^2 \hat{z}_{1i}. \tag{69}$$

Based on Lemmas 2 and 3, a pole of the zeta function $\zeta_2^{(1)}(z)$ provides upper bounds of the free energy;

$$\zeta_2^{(1)}(z) = \int_{W_2^{(1)}} \Big\{ \hat{d}_{11}^4 f_6^{(1)}(\hat{\omega}) \Big\}^z \varphi(\Phi_2^{(1)}(\hat{\omega}))|\Phi_2^{(1)}| d\hat{\omega}, \tag{70}$$

where $f_6^{(1)}(\hat{\omega})$ is a polynomial of $\hat{\omega}$. Because $|\Phi_2^{(1)}| = \hat{d}_{11}^{3q^2+2q-1}$, the zeta function $\zeta_2^{(1)}(z)$ has a pole at $z = -3q^2/4 - q/2$.

Comparing two cases $\Phi_1^{(1)}$ and $\Phi_2^{(1)}$, we obtain the upper bounds. When $p = 3q/2 + 1$, the poles are at the same position. Therefore, its order is two, which completes the proof. **(End of Proof)**

4.3 Proof of Theorem 2

The structure of the proof is the same as that of Theorem 1.

According to $P_{1|0} = 0$ and Eqs. (22), (23) and (25), there are matrices $Q_{it}^{(2)}$ and $Q_{it}^{(3)}$ consisting of ω such that

$$P_{t|t-1} = AQ_t^{(1)}A^\top + DD^\top \tag{71}$$
$$= \sum_i Q_{it}^{(2)} DD^\top Q_{it}^{(3)}. \tag{72}$$

By using $z_1 = 0$, $P_{t-1|t-2}$ and Eqs. (21), (23) and (24),

$$\begin{aligned} \hat{z}_{t|t-1} =& A\{(I - P_{t-1|t-2}C^\top(I + CP_{t-1|t-2}C^\top)^{-1}C)\hat{z}_{t-1|t-2} \\ &+ P_{t-1|t-2}C^\top(I + CP_{t-1|t-2}C^\top)^{-1}x_{t-1}\}, \end{aligned} \tag{73}$$
$$= A \sum_i Q_{it}^{(4)} DD^\top Q_{it}^{(5)} C^\top y_t^{(2)}, \tag{74}$$

where $Q_{it}^{(4)}$ and $Q_{it}^{(5)}$ are matrices consisting of ω, and $y_t^{(2)}$ is a q dimensional vector consisting of ω and $x_1, \ldots, x_{t-1}$. Due to the expression Eq. (27),

$$p(X|\omega) = \prod_{t=1}^{T} \mathcal{N}(x_t|\mu_t^{(2)}, \Sigma_t^{(2)}), \tag{75}$$
$$\mu_t^{(2)} = CA \sum_i Q_{it}^{(4)} DD^\top Q_{it}^{(5)} C^\top y_t^{(2)}, \tag{76}$$
$$\Sigma_t^{(2)} = I + C \sum_i Q_{it}^{(2)} DD^\top Q_{it}^{(3)} C^\top. \tag{77}$$

Then, the Kullback information Eq. (12) is written as

$$H(\omega) = \sum_{t=1}^{T} E\Big[\ln \frac{\mathcal{N}(x_t|0, I)}{\mathcal{N}(x_t|\mu_t^{(2)}, \Sigma_t^{(2)})} \Big]. \tag{78}$$

Let us restrict the parameter areas to $W_1^{(2)}$ and $W_2^{(2)}$ such that

$$W_1^{(2)} = \{\omega : \|c_{ij}\| < \epsilon^{(2)}\} \tag{79}$$
$$W_2^{(2)} = \{\omega : \|d_{ij}\| < \epsilon^{(2)}\}, \tag{80}$$

where $\epsilon^{(2)}$ is a small positive constant. In the similar way to the proof of Theorem 1, there is a positive constant $\alpha_1^{(2)}$ such that

$$H(\omega) \leq \alpha_1^{(2)} H_u^{(2)}(\omega) \tag{81}$$
$$H_u^{(2)}(\omega) = \sum_{t=1}^{T}\Big[\frac{1}{4}(C\sum_i Q_{it}^{(2)} DD^\top Q_{it}^{(3)} C^\top)^2 + \frac{1}{2}\mathrm{Tr} R_t^{(2)\top}\Sigma_t^{(2)-1} R_t^{(2)} S_t^{(2)}\Big], \tag{82}$$
$$R_t^{(2)} = CA\sum_i Q_{it}^{(4)} DD^\top Q_{it}^{(5)} C^\top, \tag{83}$$

where $S_t^{(2)} = E[y^{(2)} y^{(2)\top}]$, in the area $W^{(2)}$. The function $H_u^{(2)}(\omega)$ consists of terms $c_{i_1 i_5} c_{i_2 i_6} c_{i_3 i_7} c_{i_4 i_8}$ or $d_{i_1 i_2} d_{i_3 i_4} d_{i_5 i_6} d_{i_7 i_8}$. In the restricted areas $W_1^{(2)}$ and $W_2^{(2)}$, it holds that

$$H(\omega) \leq \sum_{i1,\ldots,i_4=1}^{p}\sum_{i_5,\ldots,i_8=1}^{q} c_{i_1 i_5} c_{i_2 i_6} c_{i_3 i_7} c_{i_4 i_8} f_{1 i_1 \ldots i_8}^{(2)}(\omega), \tag{84}$$
$$H(\omega) \leq \sum_{i_1,\ldots,i_8=1}^{q} d_{i_1 i_2} d_{i_3 i_4} d_{i_5 i_6} d_{i_7 i_8} f_{2 i_1 \ldots i_8}^{(2)}(\omega), \tag{85}$$

respectively, where $f_{1 i_1 \ldots i_8}^{(2)}(\omega)$ and $f_{2 i_1 \ldots i_8}^{(2)}(\omega)$ are polynomials of ω.

For Eq. (84), we can find the following blow-up $\omega = \Phi_1^{(2)}(\hat{\omega})$;

$$c_{11} = \hat{c}_{11}, \tag{86}$$
$$c_{ij} = \hat{c}_{11}\hat{c}_{ij} \ \text{(others)}. \tag{87}$$

Based on Lemmas 2 and 3, a pole of the zeta function $\zeta_1^{(2)}(z)$ provides upper bounds of the free energy;

$$\zeta_1^{(2)}(z) = \int_{W_1^{(2)}} \Big\{\hat{c}_{11}^4 f_3^{(2)}(\hat{\omega})\Big\}^z \varphi(\Phi_1^{(2)}(\hat{\omega}))|\Phi_1^{(2)}| d\hat{\omega}, \tag{88}$$

where $f_3^{(2)}(\hat{\omega})$ is a polynomial of $\hat{\omega}$. Because $|\Phi_1^{(2)}| = \hat{c}_{11}^{pq-1}$, the zeta function $\zeta_1^{(2)}$ has a pole at $z = -pq/4$.

For Eq. (85), we can find the following blow-up $\omega = \Phi_2^{(2)}(\hat{\omega})$;

$$d_{11} = \hat{d}_{11}, \tag{89}$$
$$d_{ij} = \hat{d}_{11}\hat{d}_{ij} \ \text{(others)}. \tag{90}$$

Based on Lemmas 2 and 3, a pole of the zeta function $\zeta_2^{(2)}(z)$ provides upper bounds of the free energy;

$$\zeta_2^{(2)}(z) = \int_{W_2^{(2)}} \Big\{\hat{d}_{11}^4 f_4^{(2)}(\hat{\omega})\Big\}^z \varphi(\Phi_2^{(2)}(\hat{\omega}))|\Phi_2^{(2)}| d\hat{\omega}, \tag{91}$$

where $f_4^{(2)}(\hat{\omega})$ is a polynomial of $\hat{\omega}$. Because $|\Phi_2^{(2)}| = \hat{d}_{11}^{q^2-1}$, the zeta function $\zeta_2^{(2)}(z)$ has a pole at $z = -q^2/4$.

Comparing two cases $\Phi_1^{(2)}$ and $\Phi_2^{(2)}$, we obtain the upper bounds. When $p = q$, the poles are at the same position. Therefore, its order is two, which completes the proof. **(End of Proof)**

4.4 Proof of Corollary 1

Because all matrices A, C and D are scalar, the parameters ω are denoted by a, c and d, respectively.

According to $P_{1|0} = 0$ and Eqs. (22), (23) and (25), it holds that

$$P_{t|t-1} = a^2(1 - c^2 P_{t-1|t-2}(1 + c^2 P_{t-1|t-2})^{-1})P_{t-1|t-2} + d^2 = d^2 Q_t^{(6)}, \tag{92}$$

where $Q_t^{(6)}$ is defined by the following recurrence expression;

$$Q_1^{(6)} = 0 \tag{93}$$
$$Q_{t+1}^{(6)} = 1 + \frac{a^2 Q_t^{(6)}}{1 + c^2 d^2 Q_t^{(6)}}. \tag{94}$$

By using $z_1 = 0$, $P_{t-1|t-2}$ and Eqs. (21), (23) and (24),

$$\hat{z}_{t|t-1} = a((1 - c^2 P_{t-1|t-2}(1 + c^2 P_{t-1|t-2})^{-1})\hat{z}_{t-1|t-2} + cP_{t-1|t-2}(1 + c^2 P_{t-1|t-2})^{-1} x_{t-1}), \tag{95}$$
$$= cd^2 y_t^{(3)}, \tag{96}$$

where $y_t^{(3)}$ is defined by the following expression;

$$y_1^{(3)} = 0, \tag{97}$$
$$y_{t+1}^{(3)} = \frac{a}{1 + c^2 d^2 Q_t^{(6)}}(y_t^{(3)} + Q_t^{(6)} x_t). \tag{98}$$

Due to the expression Eq. (27),

$$p(X|\omega) = \prod_{t=1}^{T} \mathcal{N}(x_t | c^2 d^2 y_t^{(3)}, 1 + c^2 d^2 Q_t^{(6)}). \tag{99}$$

Then, the Kullback information Eq. (12) is written as

$$H(\omega) = \sum_{t=1}^{T} E\Big[\ln \frac{\mathcal{N}(x_t|0,1)}{\mathcal{N}(x_t | c^2 d^2 y_t^{(3)}, 1 + c^2 d^2 Q_t^{(6)})}\Big] = \frac{1}{2}\sum_{t=1}^{T}\Big[-1 + \frac{c^4 d^4 S_t^{(3)}}{1 + c^2 d^2 Q_t^{(6)}} + \frac{1}{1 + c^2 d^2 Q_t^{(6)}} + \ln(1 + c^2 d^2 Q_t^{(6)})\Big], \tag{100}$$

where $S_t^{(3)} = E[y_t^{(3)2}]$. By assuming a restricted parameter area $\{\omega : \|c\| < \epsilon^{(3)} \text{ and } \|d\| < \epsilon^{(3)}\}$, where $\epsilon^{(3)}$ is a small positive constant, it holds that $|c^2 d^2 Q_t^{(6)}| < 1$. Then,

$$H(\omega) = \frac{c^4 d^4}{2}\sum_{t=1}^{T}\sum_{j=0}^{\infty}\Big(S_t^{(3)} + \frac{j+1}{j+2} Q_t^{(6)2}\Big)(-c^2 d^2 Q_t^{(6)})^j \tag{101}$$

on the basis of the Taylor expansion on $\ln(1+x)$ and $1/(1+x)$ for $|x| < 1$. The smallest degree with respect to ω is c^4d^4 because of the definitions of $S_t^{(3)}$ and $Q_t^{(6)}$. Then, there are positive constants $\alpha_1^{(3)}$ and $\alpha_2^{(3)}$ such that

$$\alpha_1^{(3)}c^4d^4 \leq H(\omega) \leq \alpha_2^{(3)}c^4d^4. \quad (102)$$

Based on Lemma 3, the following zeta function $\zeta^{(3)}(z)$ has the same pole and order as $\zeta(z)$,

$$\zeta^{(3)}(z) = \int c^4d^4\varphi(\omega)d\omega, \quad (103)$$

which immediately gives a pole $z = -1/4$ with $m = 2$. The learning model attains the true model in the parameter set $\{c = 0\} \cup \{d = 0\}$. The area $\{c = 0\} \cap \{d \neq 0\}$ has a pole $z = -1/4$ with $m = 1$. The area $\{c \neq 0\} \cap \{d = 0\}$ has the same pole. Therefore, the intersection point $\{c = 0\} \cap \{d = 0\}$ has the largest pole $\lambda_1 = 1/4$ with $m_1 = 2$. **(End of Proof)**

5. Discussions

First, let us compare the results of the present paper with those of [18]. The model in the previous study is limited to one-dimensional hidden state and observable variables, and the upper bounds of the coefficients are obtained as $\lambda = 1/2, m = 2$ for $z_1 = 0$ and $\lambda = 1/2, m = 1$ for $z_1 \neq 0$. The present paper provides the upper bounds in general models. The result of the corresponding dimension in Corollary 1 derives the largest pole $\lambda_1 = 1/4$ and m_1=2. The structure of singularities depends on the true parameter set $W_t = \{\omega^* : H(\omega^*) = 0\}$ and the expression of $H(\omega)$ as a polynomial with respect to ω in the neighborhood of W_t. Based on the proof of the corollary, the structure is rigorously revealed, which could not be attained in the previous study. More mathematically, the present paper proves that the function is quartic, where the dominant term is c^4d^4, while the previous study found that it is bounded by a quadratic function such as c^2d^2. This precise expression improves the asymptotic forms because the exponential part of the term c^4d^4 directly appears in the denominator of λ_1. The main results provide not only theoretical but also practical insights for the system identification and the parameter learning because the structure crucially affects the accuracy of constructing the posterior and its convergence.

Second, let us consider the relation between the posterior distribution and the asymptotic free energy. The main formula II in the book [25] shows that the asymptotic form of $-\ln Z(X^n)$ is similar to that of $F(n)$;

$$-\ln Z(X^n) + \sum_{i=1}^{n} \ln q(X_i) = \lambda_1 \ln n - (m_1 - 1)\ln\ln n + O_p(1). \quad (104)$$

According to Eqs. (1)–(2), the posterior is expressed as

$$p(\omega|X^n) = \frac{e^{-nH_n(\omega)}\varphi(\omega)}{\alpha_p n^{-\lambda_1}(\ln n)^{m_1-1}}, \quad (105)$$

$$H_n(\omega) = \frac{1}{n}\sum_{i=1}^{n} \ln \frac{q(X_i)}{p(X_i|\omega)}, \quad (106)$$

where α_p is a positive random variable satisfying the convergence in probability $\alpha_p \to 1$. The derivation of λ_1 and m_1 gives the shape of the posterior due to $\int p(\omega|X^n)d\omega = 1$. For example, let us assume that these coefficients are obtained from restricted parameter area W_r around $\{C = 0\}$ in the zeta function and there is a pole $z = -\lambda(< -\lambda_1)$ in the area W_r' around $\{A = 0\} \cap \{D = 0\}$. This indicates $\int_{W_r} p(\omega|X^n)d\omega \to 1$ for $n \to \infty$ because the restricted integral $\int_{W_r} \exp\{-nH_n(\omega)\}\varphi(\omega)d\omega$ achieves the value of the denominator. In the area W_r', on the other hand, the restricted integral converges to zero;

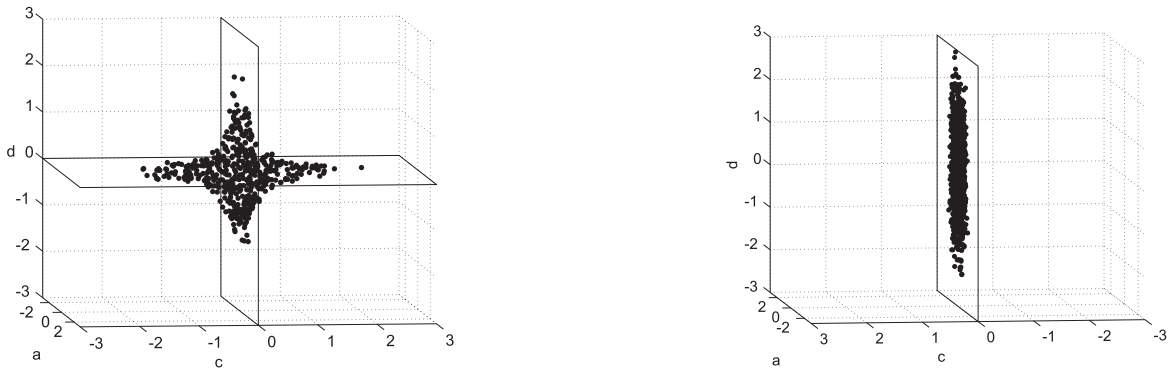
$$\int_{W_r'} \frac{e^{-nH_n(\omega)}\varphi(\omega)}{\alpha_p n^{-\lambda_1}(\ln n)^{m_1-1}}d\omega = \frac{\alpha_p' n^{-\lambda}(\ln n)^{m-1}}{\alpha_p n^{-\lambda_1}(\ln n)^{m_1-1}} = O(n^{-\lambda+\lambda_1}(\ln n)^{m-m_1}) \to 0, \quad (107)$$

where α_p' is a positive random variable satisfying the convergence in probability $\alpha_p' \to 1$. Thus, the posterior converges to the neighborhood of $\{C = 0\}$. Using this fact, we can analyze the convergence area of the posterior distribution on the basis of the restricted area providing the pole.

To attain the true model described by Eq. (29), the learning model defined by Eqs.(19) and (20) must have the true parameters in a set $\{A = 0 \cap D = 0 \cap z_1 = 0\} \cup \{C = 0\}$. Figure 1 borrowed from [18] describes experimental sampling from the posterior distributions in one-dimensional cases, where the parameter space is expressed as (a, c, d, z_1). For simplicity, Figure 1 shows the space (a, c, d). The vertical and horizontal planes indicate $\{c = 0\}$ and $\{d = 0\}$, respectively. The sampling method is the Markov chain Monte Carlo (MCMC) method [26], with 1,000 training sequences and $T = 10$. The left and right panels show cases of $z_1 = 0$ and $z_1 \neq 0$, respectively.

In $z_t = 0$, the points are located around the subspace $\{c = 0\} \cup \{d = 0\}$, for which the parameters express the true model. Based on the relation between $F(n)$ and $p(\omega|X^n)$, the proof of Corollary 1 shows that the sampled points will converge to the origin because the area, where both c and d are close to zero, provides the pole $z = -1/4$. As shown in the figure, the points are still scattered along the subspace, which implies that the sampling does not precisely construct the posterior. In singular models, the posterior converging to a point generally requires accurate sampling techniques. The figure confirms this difficulty.

In $z_t \neq 0$, the points are around the subspace $\{c = 0\}$. Theorem 1 shows the bounds with coefficient $\lambda = 1/2$ because $p < 3q/2 + 1$ for one-dimensional case. This coefficient is obtained by the neighborhood of $\{C = 0\}$. Therefore, the plot supports the validity of the bounds. Theorem 1 also implies that the posterior distribution can converge to the neighborhood of $\{A = 0\} \cap \{D = 0\} \cap \{z_1 = 0\}$ when $p > 3q/2 + 1$ in higher dimensional cases.



(a) $z_1 = 0$ (b) $z_1 \neq 0$

Fig. 1 Examples of the posteriors borrowed from [18].

Third, we elucidate importance of the simple setting, where the true model is $\mathcal{N}(x_t|0, I)$. Assume that the hidden state has a general dimension; $z_t \in \mathbf{R}^{q^*}$ in the true model, and $z_t \in \mathbf{R}^{q^*+q}$ in a learning model. According to analyses of other singular models e.g. [24], [25], the coefficient λ_1 is decomposed as $\lambda_1 = d_t/2 + \lambda_r$, where d_t is the essential parameter dimension of the true model and λ_r is the remaining part. The value of this remaining part is affected by redundant hidden states of the learning model. The structure of singularities found in the present paper exists in a parameter space of the general case, which means that the calculation of λ_r requires similar derivation to the proofs in Sect. 4. Therefore, analysis of the true model $\mathcal{N}(x_t|0, I)$ is mathematically important though singularities are more complicated in practical cases.

Last, let us discuss the hidden state estimation from the point of view of the posterior distribution. In the practical application, the state estimation with the Kalman filter is based on the optimal parameter, i.e. the state space equations with the optimal coefficients are given. To find the coefficients, the maximum likelihood method and the MAP method are employed as the parameter search. The posterior distribution provides a possible estimator in these methods. For example, one of the points in Fig. 1 will be selected as the MAP estimator. Thus, observation of the posterior distribution enables us to predict the result of the hidden state estimation.

As confirmed in one-dimensional case, Theorem 1 shows that the posterior distribution can be located around the area $\{C = 0\}$ for $p < 3q/2 + 1$. It should be reminded that the estimated values of C are not exactly zero due to the noises. Then, a hidden state estimation result shows that z_t has a movement following Eq. (19) and weakly influences x_t as showed in Eq. (20) though the true model does not have z_t. For the same reason, the random walk of z_t will be estimated due to the posterior distribution around $\{C = 0\} \cap \{D = 0\}$ in Theorem 2. This adverse estimation will occur when the learning model has redundant dimension of z_t in higher dimensional models of practical situations.

6. Conclusions

The present paper extended the previous study on LDS [18] and more precisely analyzed singularities in the parameter space. Due to the structure of singularities, the asymptotic forms of the marginal likelihood and the generalization error are clarified, which provides theoretical and practical insights for the system identification. The previous study pointed out that the singularities adversely affect hidden state estimation based on the Kalman filter when LDS has one-dimensional variables. The results of the present paper indicate that general models also have the same effect. To prevent the adverse estimation, the dimension of hidden state variables must be carefully detected. The structure of singularities analyzed in the present paper will be helpful to leverage the criteria for singular models. The criteria such as [22], [23] require the structure of singularities, where the true model has the hidden states $\dim z_t = q^* > 0$. Unfortunately, the results of the present paper focus on the case $q^* = 0$, which is not enough to directly apply to the model selection. Our calculation on the zeta function, however, will provide fundamental knowledge for the general true model, and extending it to the general case $q^* > 0$ is one of our important future works.

Acknowledgment

This research was partially supported by the Kayamori Foundation of Informational Science Advancement and KAKENHI 23500172.

References

[1] D. Obradovic, H. Lenz, and M. Schupfner, "Sensor fusion in siemens car navigation system," Proc. MLSP 2004, pp.655–664, 2004.
[2] N. Funk, "A study of the Kalman filter applied to visual tracking," Tech. Rep. Project for CMPUT 652, University of Alberta, 2003.
[3] F. Sahba, H.R. Tizhoosh, and M.M. Salama, "A coarse-to-fine approach to prostate boundary segmentation in ultrasound images.," Biomed Eng Online, vol.4, p.58, 2005.
[4] R.E. Kalman, "A new approach to linear filtering and prediction

problems," J. Basic Engineering, vol.82, pp.35–45, 1960.
[5] S. Watanabe, "Algebraic analysis for nonidentifiable learning machines," Neural Comput., vol.13, no.4, pp.899–933, 2001.
[6] S. Watanabe, "Algebraic geometrical methods for hierarchical learning machines," Neural Netw., vol.14, no.8, pp.1049–1060, 2001.
[7] K. Yamazaki, M. Aoyagi, and S. Watanabe, "Asymptotic analysis of Bayesian generalization error with Newton diagram," Neural Netw., vol.23, pp.35–43, January 2010.
[8] S. Lin, "Asymptotic approximation of marginal likelihood integrals," arXiv:1003.5338, March 2010.
[9] K. Yamazaki and S. Watanabe, "Singularities in mixture models and upper bounds of stochastic complexity," Int. J. Neural Netw., vol.16, pp.1029–1038, 2003.
[10] K. Yamazaki and S. Watanabe, "Stochastic complexity of bayesian networks," Proc. UAI, pp.592–599, 2003.
[11] K. Yamazaki and S. Watanabe, "Singularities in complete bipartite graph-type boltzmann machines and upper bounds of stochastic complexities," IEEE Trans. Neural Netw., vol.16, no.2, pp.312–324, 2005.
[12] K. Yamazaki and S. Watanabe, "Generalization errors in estimation of stochastic context-free grammar," The IASTED International Conference on ASC, pp.183–188, 2005.
[13] K. Yamazaki and S. Watanabe, "Algebraic geometry and stochastic complexity of hidden Markov models," Neurocomputing, vol.69, no.1-3, pp.62–84, 2005.
[14] M. Aoyagi and S. Watanabe, "Stochastic complexities of reduced rank regression in bayesian estimation," Neural Netw., vol.18, pp.924–933, 2005.
[15] D. Rusakov and D. Geiger, "Asymptotic model selection for naive bayesian networks," J. Machine Learning Research, vol.6, pp.1–35, 2005.
[16] M. Aoyagi, "Stochastic complexity and generalization error of a restricted boltzmann machine in bayesian estimation," J. Mach. Learn. Res., vol.11, pp.1243–1272, 2010.
[17] P. Zwiernik, "An asymptotic behaviour of the marginal likelihood for general markov models," J. Mach. Learn. Res., vol.12, pp.3283–3310, Nov. 2011.
[18] T. Naito and K. Yamazaki, "A study on bayesian learning of one-dimensional linear dynamical systems," Neural Information Processing, ed. C.S. Leung, M. Lee, and J.H. Chan, Lecture Notes Comput. Sci., vol.5863, pp.110–117, Springer, 2009.
[19] H. Akaike, "A new look at the statistical model identification," IEEE Trans. Autom. Control, vol.19, pp.716–723, 1974.
[20] G. Schwarz, "Estimating the dimension of a model," Annals of Statistics, vol.6, no.2, pp.461–464, 1978.
[21] S. Watanabe, "Equations of states in singular statistical estimation," Neural Netw., vol.23, no.1, pp.20–34, 2010.
[22] K. Yamazaki, K. Nagata, and S. Watanabe, "A new method of model selection based on learning coefficient," Proc. International Symposium on Nonlinear Theory and its Applications, pp.389–392, 2005.
[23] K. Yamazaki, K. Nagata, S. Watanabe, and K.R. Müller, "A model selection method based on bound of learning coefficient," LNCS, pp.371–380, Springer, 2006.
[24] M. Aoyagi and S. Watanabe, "Stochastic complexities of reduced rank regression in Bayesian estimation," Neural Networks, vol.18, pp.924–933, 2005.
[25] S. Watanabe, Algebraic Geometry and Statistical Learning Theory, Cambridge University Press, New York, NY, USA, 2009.
[26] C. Andrieu, N. de Freitas, A. Doucet, and M.I. Jordan, "An introduction to MCMC for machine learning," Machine Learning, vol.50, no.1-2, pp.5–43, 2003.

Takuto Naito received the B.E. and M.E. degrees in information engineering from Tokyo Institute of Technology in 2008 and 2010, respectively. In 2010, he joined IT Products Business Division, AVC Networks Company, Panasonic Corporation, where he has been engaged in development of personal computers.

Keisuke Yamazaki received B.E., M.E., and Ph.D. degree in information engineering from Tokyo Institute of Technology in 2001, 2002, and 2004, respectively. He is now an assistant professor in Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology. His research interests include information science and machine learning.