情報論的学習理論テクニカルレポート 2009
Technical Report on Information-Based Induction Sciences 2009 (IBIS2009)

1次元線形動的システムの特異性とベイズ汎化誤差への影響

Singularities of One-dimensional Linear Dynamical Systems and its Effect on the Bayesian Generalization Error

内藤 卓人* 山崎 啓介†
Takuto Naito Keisuke Yamazaki

Abstract: Linear dynamical systems are widely used in such fields as system control and time-dependent data analysis. Such a system can be regarded as a statistical parametric model, where the coefficients of the state space equations are unknown and given as parameters. The properties of parameter learning have not yet been established, in spite of a wide range of applications. Therefore, this paper investigates the system from the viewpoint of learning theory. It is revealed that the system has singularities in the parameter space. The generalization error measured by the prediction accuracy for unseen data sequences is reduced, due to the presence of these singularities.

Keywords: Kalman Filter, Bayesian Learning, Time-Series Data Analysis

1 Introduction

Linear dynamical systems are widely used for modeling practical complex systems with hidden variables such as object tracking in image processing [4], and position detection in car navigation systems [6]. The system is described via state space equations containing both observable and hidden variables. The Kalman filter [5] is an algorithm to estimate the hidden variables from coefficients given preliminarily .

It is important to be able to estimate coefficients using the observable data when the coefficients are unknown. The system is regarded as a parametric learning model, in which the coefficients correspond to parameters. As seen in Section 2, the system is expressed as a generative probability model of the data because the process and observation noises are taken into account.

*東京工業大学, 知能システム科学専攻, 226-8503 神奈川県横浜市緑区長津田町 4259 R2-5, tel. 045-924-5018, e-mail naitaku@cs.pi.titech.ac.jp,
Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, R2-5, 4259 Nagatsuta, Midori-Ku, Yokohama, Kanagawa
†東京工業大学, 精密工学研究所, 226-8503 神奈川県横浜市緑区長津田町 4259 R2-5, tel. 045-924-5018, e-mail k-yam@pi.titech.ac.jp, Precision and Intelligence Laboratory, Tokyo Institute of Technology, R2-5, 4259 Nagatsuta, Midori-Ku, Yokohama, Kanagawa

Parametric models generally fall into two types, *regular* and *singular*. If the relation between the parameter and the expressed probability function is one-to-one, a model is referred to as regular. Otherwise, it is singular. Therefore, a singular model has a set of parameters indicating the same function, in which there are singularities. Because of the singularities, conventional analysis is not applicable; model selection criteria for regular models such as AIC [1] and BIC [7] are inappropriate. An algebraic geometrical method has been developed for Bayesian learning to reveal the asymptotic generalization error and the marginal likelihood for several singular models [8]. According to its application to several models, the presence of singularities results in unique properties of the learning process [3, 9].

In spite of a wide range of applications, properties of a linear dynamical system are still unknown in terms of a learning model. Therefore, the present paper investigates such a system both theoretically and experimentally. We confirm that the system is a singular model and analyze the Bayesian generalization error based on the algebraic geometrical method. Here, the error is defined as the prediction accuracy for unseen time-sequence data. This *prediction* is different from

that of the conventional Kalman situation in which the primary concern is the set of hidden variables rather than the observable sequences. Nevertheless, our analysis can also provide an insight into hidden variable estimation.

The remainder of the paper is organized as follows. Section 2 formulates the system. Section 3 introduces Bayesian learning and summarizes the algebraic geometrical method. Section 4 contains our main contributions, deriving a theoretical upper bound of the generalization error and showing experimental results for the error. Section 5 contains a discussion and our conclusions.

2 Linear Dynamical Systems

Linear dynamical systems can be described by state space models with hidden state variables:

$$z_{t+1} = Az_t + Dw_t, \tag{1}$$
$$x_t = Cz_t + v_t, \tag{2}$$

where $z_t \in \mathbf{R}^q$ is the hidden state vector at time t, $x_t \in \mathbf{R}^p$ is an output vector, $w_t \in \mathbf{R}^q$ and $v_t \in \mathbf{R}^p$ are process and observation noises, respectively. These noises are assumed follow a standard normal distribution. $A \in \mathbf{R}^{q\times q}$ is the state matrix , $C \in \mathbf{R}^{p\times q}$ is the output matrix and the elements of $D \in \mathbf{R}^{q\times q}$ are the coefficients of the process noise.

The Kalman filter is known as an efficient recursive filter that estimates hidden states from a series of outputs. In what follows, the notations $\hat{z}_{n|m}$ and $P_{n|m}$ represent the estimates of z at time n and its error covariance matrix, respectively, when observations from $t = 1$ to $t = m$ are given. The Kalman filter has two phases: **Predict** and **Update**. The algorithms are described as follows:

Predict

$$\hat{z}_{t|t-1} = A\hat{z}_{t-1|t-1} \tag{3}$$
$$P_{t|t-1} = AP_{t-1|t-1}A^\top + DD^\top \tag{4}$$

Update

$$K_t = P_{t|t-1}C^\top \left(I + CP_{t|t-1}C^\top\right)^{-1} \tag{5}$$
$$\hat{z}_{t|t} = \hat{z}_{t|t-1} + K_t\left(x_t - C\hat{z}_{t|t-1}\right) \tag{6}$$
$$P_{t|t} = \left(I - K_tC\right)P_{t|t-1} \tag{7}$$

where I is a unit matrix and K_t is called the Kalman gain. Firstly, the current state z_t is estimated as $\hat{z}_{t|t-1}$ from the estimated state of the previous time $t-1$ (Eq.3). Then, a more refined value for $\hat{z}_{t|t}$ is calculated on the basis of $\hat{z}_{t|t-1}$ after an observation x_t is provided (Eq.6).

From the viewpoint of machine learning, a linear dynamical system can be regarded as a learning model whose parameters are A, C, D and z_1. The variable z_1 indicates the initial state. Let $X = (x_1, x_2, \ldots, x_T) \in \mathbf{R}^{p\times T}$ be the vector of observations. The probability $p(X|w)$, where the parameters $w = (A, C, D, z_1)$, can be calculated as follows:

$$p(X|w) = p(x_1|w)\prod_{t=2}^{T} p(x_t|x_1, \ldots, x_{t-1}, w). \tag{8}$$

Using the hidden state z_t,

$$p(x_t|x_1, \ldots, x_{t-1}, w) = \int p(x_t|z_t, w)p(z_t|x_1, \ldots, x_{t-1}, w)dz_t. \tag{9}$$

Let $\mathcal{N}(\cdot|\mu, \Sigma)$ be a multivariate normal distribution with mean μ and covariance matrix Σ. By the definition of a linear dynamical system (Eq.2) and the derivation of the Kalman filter,

$$p(x_t|z_t, w) = \mathcal{N}(x_t|Cz_t, I), \tag{10}$$
$$p(z_t|x_1, \ldots, x_{t-1}, w) = \mathcal{N}(z_t|\hat{z}_{t|t-1}, P_{t|t-1}). \tag{11}$$

Therefore, $p(x_t|x_1, \ldots, x_{t-1}, w)$ is also a normal distribution described by

$$p(x_t|x_1, \ldots, x_{t-1}, w) = \mathcal{N}(x_t|C\hat{z}_{t|t-1}, I + CP_{t|t-1}C^\top). \tag{12}$$

Eq.8 can be expressed as

$$p(X|w) = \prod_{t=1}^{T} \mathcal{N}(x_t|C\hat{z}_{t|t-1}, I + CP_{t|t-1}C^\top). \tag{13}$$

where we define $\hat{z}_{1|0} = z_1$ and $P_{1|0} = 0$.

Let $X^n = (X_1, X_2, \ldots, X_n)$ be a set of i.i.d. training samples. Each X_i is a time sequence defined by $X_i = (x_1^i, x_2^i, \ldots, x_T^i)$. The likelihood of the parameter $w = (A, C, D, z_1)$ can be calculated as

$$L(w) = \prod_{i=1}^{n} p(X_i|w) = \prod_{i=1}^{n}\prod_{t=1}^{T} \mathcal{N}(x_t^i|C\hat{z}_{t|t-1}^i, I + CP_{t|t-1}^iC^\top) \tag{14}$$

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

3 Bayesian Learning and the Generalization Error

This section describes Bayesian learning for time series data and the theoretical analysis of the generalization error.

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. 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 $p(X|w)$ be a learning model, and $\varphi(w)$ be an a priori probability distribution. The a posteriori probability distribution is defined by

$$p(w|X^n) = \frac{1}{Z(X^n)}\varphi(w)\prod_{i=1}^{n} p(X_i|w) \tag{15}$$

where $Z(X^n)$ is a normalizing constant. The Bayesian predictive distribution is defined by

$$p(X|X^n) = \int p(X|w)p(w|X^n)dw. \tag{16}$$

The Bayesian generalization error $G(n)$ is defined by

$$G(n) = E_{X^n}\Big[\int q(X)\log\frac{q(X)}{p(X|X^n)}dX\Big], \tag{17}$$

which is the average Kullback information from the true distribution to the predictive distribution.

The remainder of this section summarizes the algebraic geometrical method for deriving the asymptotic form of the error [8]. Let $H(w)$ be the Kullback information from the true distribution $q(X)$ to the learner $p(X|w)$,

$$H(w) = \int q(X)\log\frac{q(X)}{p(X|w)}dX. \tag{18}$$

The function $\zeta(z)$ of one complex variable z, defined by

$$\zeta(z) = \int H(w)^z\varphi(w)dw, \tag{19}$$

is referred to as the zeta function. It is known that this zeta function is holomorphic in the region $\mathrm{Re}(z) > 0$, and can be analytically continued to the meromorphic function on the entire complex plane. Then the poles are all real, negative and rational numbers. Let $0 > -\lambda_1 > -\lambda_2 > \ldots$ be a sequence of poles, and $m_1, m_2, \ldots$ be the respective orders. The asymptotic form of the generalization error is expressed as

$$G(n) = \frac{\lambda_1}{n} - \frac{m_1 - 1}{n\log n} + o\Big(\frac{1}{n\log n}\Big) \tag{20}$$

for $n \to \infty$. In many cases, it is not straightforward to find the largest pole $-\lambda_1$ and its order m_1 [3]. When a pole $z = -\lambda$ and its order m have been calculated, an upper bound is derived as

$$G(n) \leq \frac{\lambda}{n} - \frac{m - 1}{n\log n} + o\Big(\frac{1}{n\log n}\Big). \tag{21}$$

4 Analysis of the Generalization Error

This section analyzes the Bayesian generalization error for linear dynamical systems. In order to investigate the effect of redundant hidden states, we study an essential case, in which the learning model has a hidden variable and the true model generates i.i.d. sequences. This is the simplest setting for singularities to exist in the parameter space because the i.i.d. model can be regarded as a model with no hidden states. For simplicity, we assume that the output vector is one dimensional, where z_t, x_t, A, C, and D are all scalar. Moreover, we assume that the first hidden state is fixed as $z_1 = 0$. Formally, the learning model is defined as

$$z_{t+1} = az_t + dw_t, \tag{22}$$

$$x_t = cz_t + v_t, \tag{23}$$

where $z_t, x_t \in \mathbf{R}^1$ and w_t and v_t are distributed from $\mathcal{N}(\cdot|0, 1)$. The parameter is expressed as $w = (a, c, d)$. The true model is a one-dimensional normal distribution $\mathcal{N}(x_t|0, 1)$ for all t, i.e. $x_t = v_t$. Following Eq. 13, the true model is given by

$$q(X) = \prod_{t=1}^{T}\mathcal{N}(x_t|0, 1). \tag{24}$$

4.1 Theoretical analysis

Based on the algebraic geometrical method, the error has the following bound:

Theorem 4.1 *When the true model and a learning model are defined by Eq.24 and Eqs 22-23, respectively, the Bayesian generalization error is bounded above as follows:*

$$G(n) \leq \frac{1}{2n} - \frac{1}{n\log n} + o\Big(\frac{1}{n\log n}\Big), \tag{25}$$

where $z_1 = 0$ and the training sample size n is sufficiently large.

Sketch of Proof: Because the parameter set $\{c = 0\}$ attains $p(X|w) = q(X)$, there is a function $f_c(w)$ such that $H(w) = c^2 f_c(w)$. The set $\{d = 0\}$ ensures the same property for $H(w)$. Thus, there is a polynomial $f(w)$ such that $H(w) = c^2 d^2 f(w)$. We can find a limited support W of the parameter space, such that $H(w) \leq Cc^2 d^2$. Here C is a positive constant. Considering the following zeta function

$$\zeta_1(z) = \int_W \{Cc^2 d^2\}^z dw, \tag{26}$$

the pole $z = -\mu$ is a lower bound of $z = -\lambda_1$ [8]. We can find a pole $\mu = 1/2$ and its order $m = 2$. Combining with Eq. 21, we derive the following leading terms for the bound,

$$\frac{1}{2n} - \frac{1}{n \log n}, \tag{27}$$

which completes the proof.

End of Proof

If the initial state is unknown and is regarded as a parameter such as $w = (a, c, d, z_1)$, we can extend Theorem 4.1 as follows.

Corollary 4.1 *Under the same setting as Theorem 4.1, the error has an upper bound*

$$G(n) \leq \frac{1}{2n} + o\Big(\frac{1}{n}\Big). \tag{28}$$

We omit the proof for lack of space.

4.2 Experimental results

We experimentally evaluate whether the bound is valid when finite training data are given. Sampling from the a posteriori distribution, the predictive distribution is given by

$$p(X|X^n) \simeq \frac{1}{M} \sum_{j=1}^{M} p(X|w_j), \tag{29}$$

where $(w_1, \ldots, w_M)$ are sampled from $p(w|X^n)$. We use the Markov chain Monte Carlo (MCMC) method for the sampling technique [2]. The generalization error is approximated by

$$G(n) \simeq E_{X^n}\Big[\frac{1}{N} \sum_{i=1}^{N} \log \frac{q(X_i)}{p(X_i|X^n)}\Big]. \tag{30}$$

The experimental settings are as follows. The length of the time sequence is $T = 10$. The number of test data sequences is $N = 1,000$. The number in the MCMC sample is $M = 500$. We obtain the expectation $E_{X^n}[\cdot]$ over 100 sets of training data. The a priori distribution is a normal distribution for a, c and d.

Figure 1-(a) describes an example of sampling from the a posteriori distribution in the parameter space (a, c, d). The vertical and horizontal planes indicate $\{c = 0\}$ and $\{d = 0\}$, respectively. The points are located around the subspace $\{c = 0\} \cup \{d = 0\}$, for which the parameters express the true model.

Figure 1-(b) summarizes the error values corresponding to $n = 250, 500, 750$ and $1,000$. The horizontal and vertical axes describe the number of training data sequences and the error value, respectively. The heavy line depicts experimental values for $G(n)$. The dotted line is the upper bound of Theorem 4.1. The upper bound is valid as seen in the graph.

5 Discussions and Conclusions

First, let us discuss the upper bound of the generalization error. In the regular case, the error has the following asymptotic form,

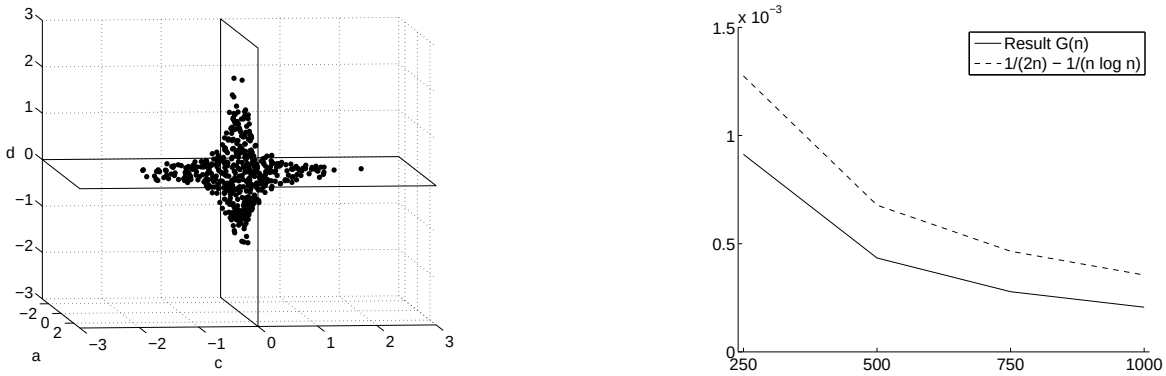
$$G(n) = \frac{\dim w}{2n} + o\Big(\frac{1}{n \log n}\Big), \tag{31}$$

which means that $\lambda_1 = \dim w/2$ and $m_1 = 1$. Note that even a singular model has this asymptotic form if the true and learning models have the same dimension of the hidden state vector. The asymptotic form indicates that the cost to fit all parameters determines the error as the dimension $\dim w$ appears. Comparing Theorem 4.1 with the regular case, we can derive the result that the error is much smaller, i.e.

$$\begin{aligned} G(n) &\leq \frac{1}{2n} - \frac{1}{n \log n} + o\Big(\frac{1}{n \log n}\Big) \\ &< \frac{3}{2n} + o\Big(\frac{1}{n \log n}\Big), \end{aligned} \tag{32}$$

which confirms that the fitting cost for redundant parameters is not strongly reflected in the error.

Thus far, we have focused on prediction of the unseen observable data sequence X. Next, we consider estimation of the hidden states z_t. According to the a posteriori distribution, there are two regions for the optimal parameters; one is around $c = 0$ and the other



(a) (b)

Figure 1: An example of the a posteriori distribution and the generalization error.

is around $d = 0$. They imply completely different behaviors of the hidden state. The former, $c = 0$, indicates that a and d can take any value, by which $q(X) = p(X|w)$. Thus, there are no constraints on the movement of the hidden state. By taking into account $z_1 = 0$, the latter, $d = 0$, contrarily implies that there is no movement because $z_t = 0$ for all times t. If several hidden variables in the true model stop moving due to disorder in a practical situation, the desired estimation is $d = 0$. However, $c = 0$ can also be an estimated result; these variables move on the basis of arbitrarily-estimated a and d. This adverse estimation can occur along any dimension of the hidden state vector. Therefore, detection of hidden variable size is an essential problem to solve.

Finally, we state our conclusions. The present paper establishes that linear dynamical systems are singular models. The singularities ensure that the upper bound of the Bayesian generalization error is small. The experimental results indicate that the bound is valid. Moreover, the a posteriori distribution implies that estimation of hidden states cannot be appropriate if there are redundant hidden variables.

Acknowledgment

This research was partially supported by the Ministry of Education, Science, Sports and Culture in Japan, Grant-in-Aid for Scientific Research 18079007.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, Vol. 19, pp. 716–723, 1974.

[2] Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, Vol. 50, No. 1-2, pp. 5–43, 2003.

[3] Miki Aoyagi and Sumio Watanabe. Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Networks*, Vol. 18, pp. 924–933, 2005.

[4] N. Funk. A study of the Kalman filter applied to visual tracking. Technical Report Project for CMPUT 652, University of Alberta, 2003.

[5] R. E. Kalman. A new approach to linear filtering and prediction problems. *J. Basic Engineering*, Vol. 82, pp. 35–45, 1960.

[6] D. Obradovic, H. Lenz, and M. Schupfner. Sensor fusion in siemens car navigation system. *Proc. of MLSP 2004*, pp. 655–664, 2004.

[7] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, Vol. 6 (2), pp. 461–464, 1978.

[8] Sumio Watanabe. Algebraic analysis for nonidentifiable learning machines. *Neural Computation*, Vol. 13, No. 4, pp. 899–933, 2001.

[9] Keisuke Yamazaki and Sumio Watanabe. Algebraic geometry and stochastic complexity of hidden Markov models. *Neurocomputing*, Vol. 69, No. 1-3, pp. 62–84, dec 2005.