博士論文（要約）

A Study on Extraction of Latent-Dynamical Structure from Time-Series Data

(時系列データからの潜在動力学構造抽出に関する研究)

村田 伸

Abstract

It was important to find the latent structure for understanding the nature of subject. In this thesis, we studied about the extraction of latent dynamical structure from the time-series data. Elucidation of the dynamical structure lead us to understand the subject. Particularly, three kinds of time-series data were studied in this thesis.

First, the associative memory model with synaptic depression was analyzed by using dimensionality reduction techniques. Although the dynamics of the model is known to be difficult to analyze theoretically, we found limit cycle like dynamics in the model. This result suggested that information processing of the model was performed through the dynamics.

Second, the coherent phonon signal was analyzed. The signal consists of sum of the damping oscillation with substance-specific normal mode frequencies. Conventional methods to analyze the coherent phonon are the Fourier transformation and the Wavelet transformation. We applied the dimensionality reduction method, instead, and showed that the method could separate the signal and background noise.

Finally, time-series spectral data was studied. The spectral data has multi-peak structure and the parameters of the peaks reflect the nature. We adopted the Bayesian inference frameworks to propose a method to analyze time-series spectral data.

Acknowledgments

This thesis summarized the works done during Ph. D. course of the Department of Complexity Science and Engineering, Graduate School of Frontier Sciences, The University of Tokyo. Many people gave me a lot of helps in explicitly or implicitly ways.

First, I would like to thank my supervisor, Prof. Masato OKADA, who provides guidance to me and give me a lot of advises over my research.

I also would like to thank the all members of my Ph. D. committee, Professor Ichiro YOSHIKAWA, Professor Akinao NOSE, Professor Kojiro SUZUKI, and Professor Akira EJIRI for giving me useful comments on my draft of this thesis.

I would like to thank all the members of the lab. Especially, I would like to thank Dr. Kenji NAGATA, who was an assistant professor, for giving me advises, and also Dr. Yosuke OTSUBO for discussing with me about my first work in my graduate school.

I also thank to the friends of mine with whom I spent great time.

Finally, I'd like to thank my family, who encourage me, support me, and let me do what I want.

Contents

Chapter 1

Introduction

These days, the performance of measurement tools is improving significantly and comes to bring us high-dimensional big data. When we attempt to handle such data, the curse of dimensionality would appear, and it might prevent us from comprehending the essential feature lying in the data or might delude us into drawing the wrong conclusion. How can we tackle those data and find the true nature? One possible way of solving this question is to detect a few essential elements to describe the phenomena and to discover a (low-dimensional) latent structure behind the phenomena. This is called a "sparse modeling" framework based on the assumption that a combination of a few elements causes a diversity expressed in the data. Another strategy for those problems is a Bayesian inference which is a probabilistic framework and make use of a prior knowledge of generation and observation.

This thesis discusses time-series data through the whole. Especially, we consider the case when the variables of interest cannot be obtained directly, as is often the case with many fields of science. Hence, it is scarcely avoided to discuss "the observation process," or relationships between the latent variables and observable ones. In measurement science, the process might be known and modeled well. We can estimate the unobservable data by utilizing the *a priori* knowledge of the process. In other fields of science, it is more often cases, however, that we do not know the relationships. In this case, we need some principal, assumption, or strategy to tackle the problem. It depends on the case, of course, but we expect to compress the data to be one of the answers. Moreover, other questions we should consider are left; what is "the latent structure" of time-series and what is suitable for the building blocks for dynamics.

One simple candidate is the linearization of the system around a fixed point. However the dynamics express nonlinearity noticeably, linearization of the dynamics can be guaranteed around the fixed points. Once we find the linearized dynamics, the dynamics can be expressed by the eigenvalues and eigenvectors gov-

erning the dynamics. Figure 1.1 shows schematic figures. For the discrete time

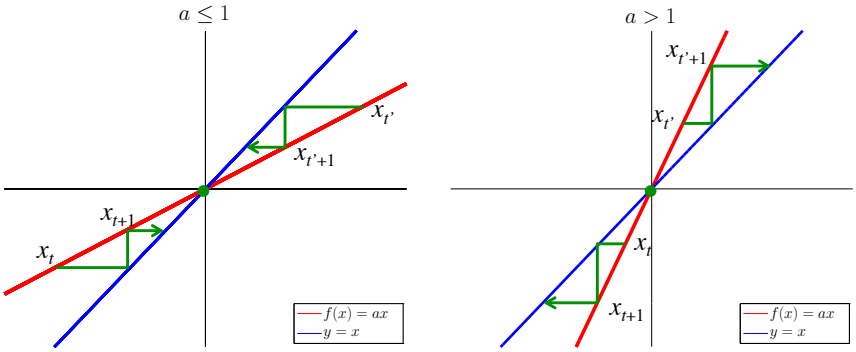


Figure 1.1: Near a fixed point dynamics.

system, when the gradient at the fixed point $a \leq 1$ then the system moves toward the fixed point, on the other hand, it goes away. There are many attractor in the nonlinear dynamical system, and hence, it is important to understand such dynamics locally. Hereby, we mainly focus on (quasi-)equilibria of the system.

1.1 State space modeling

A state space modeling is utilized for time-series data analyses in many scientific or engineering fields [1, 2]. Figure 1.2 shows a typical example of the modeling. It is assumed that there are two classes of variables: latent variables and observable ones, and that two types of relationship are found among the variables: dynamics and observation.

The latent variables, as the name suggests, cannot be observed directly, and a dynamical structure is assumed to be existing among them:

$$x_t = g(x_{t-1}; \theta).$$

Here, parameters θ are introduced to the process in order to specify the latent dynamics. The relationship, namely $g(\cdot)$, can be linear or nonlinear, and deterministic or stochastic. Note that, while the number of the latent variables is at least $\mathcal{O}(T)$ with respect to the number of time steps T the number of the parameters is $\mathcal{O}(1)$. Thus, θ itself is the latent dynamical structure which we try to extract from the data.

For instance, the left side of Fig. 1.3 shows a time-series data generated by a stochastic process. As you can see, the time-series data might have correlation with the past data of their own. When the current value, x_t, and the value just before, x_{t-1}, are plotted on the vertical and horizontal axis, respectively, the clear

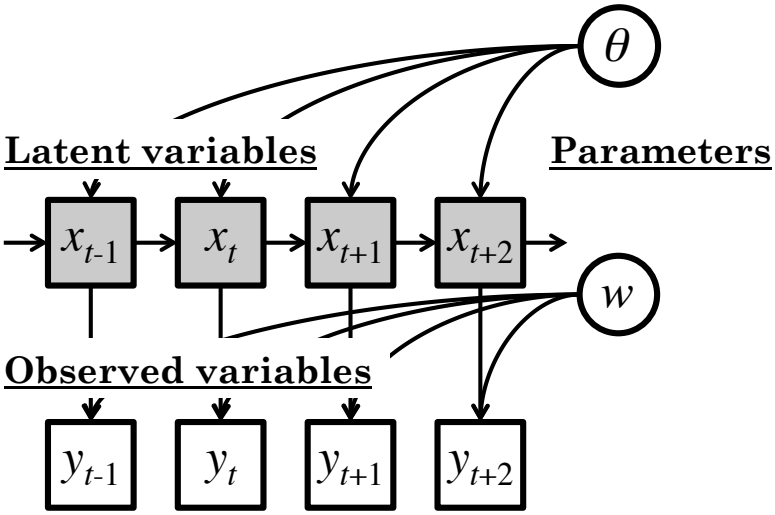


Figure 1.2: Concept of the state space modeling.

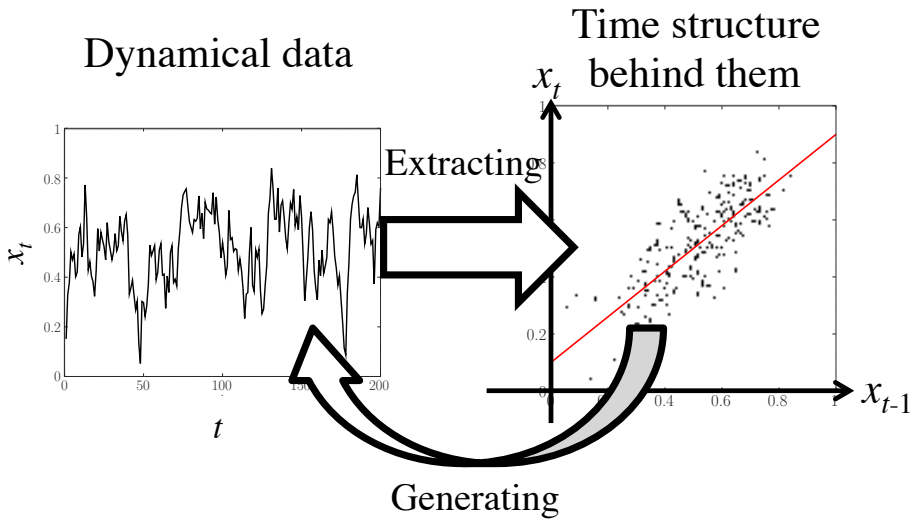


Figure 1.3: An example of a dynamical process and its latent structure.

linear correlation can be found (the right side of Fig. 1.3). This is because that the stochastic process generating the time-series is a autoregressive process of the form, $x_t = \theta_1 x_{t-1} + \theta_2 +$ noise shown as the red line in Fig. 1.3. In this example, we can see that the dynamics can be described only by the two parameters; the gradient θ_1 and the intercept θ_2. Hence, our goal is to find such parameters from the data.

Recall that we only obtain the observed variables rather than the latent variables exhibiting the dynamical properties. In this case, a problem setting becomes more complicated, that is, we need to consider "an observation process" between the latent and observable variables:

$$y_t = f(x_t;\, w).$$

Here, w is a parameter related with the observation process. The observation process can also be linear or nonlinear, and deterministic or stochastic as well as the dynamical process. Although the observable variables only depends on the latent variables, the dynamical structure of latent variables causes apparent dynamical structure among the observable data.

1.2 Related works

Now, we need to find (i) the latent variables x_t, (ii) the parameter w related to observation, (iii) the parameter θ governing the latent dynamics, and (iv) the relations $f(\cdot)$ and $g(\cdot)$ from the observation y_t. There are many algorithms and frameworks for time-series analyses. Thus, we will briefly summarize them and point out what our thesis aims at.

1.2.1 Kalman filter and its variant

The Kalman filter [1] is a widely applied algorithm in the engineering field. Note that, the Kalman filter assumes that the observation and dynamical process are already detected, hence, it is intended to estimate the hidden time-series. This algorithm repeats the prediction of the hidden state and the update the state from the observation. Under certain conditions, the Kalman filter is a optimal estimator of the latent time-series.

Since linear dynamics and linear observations are assumed for original Kalman filter, there are some extensions of the method to nonlinear systems; such as extended Kalman filter [3], unscented Kalman filter [4] and particle filter [5]. They approximate nonlinear function itself, mean and covariance, or conditional expectation.

Although, the Kalman filter and its variants are widely utilized in many fields, they are not suitable for *the extraction of the latent dynamical structure* discussed in this thesis. We focus on the situation in which the model is not perfectly determined.

1.2.2 Bayesian inference

Bayesian inference is a probabilistic framework for estimating the parameters from the observed data. The generating and observation processes are modeled stochastically, and a prior knowledge can be naturally introduced to the model. Hence, this frameworks are very useful in natural science fields, especially in measurement science. For example, Doppler effect of the light brings a Gaussian shape broadening to the spectral. In such case the observation process can be reduced to the nonlinear fitting problem. Thus, we can tackle the nonlinearity of observation process based on the stochastic inference. This stochastic frameworks assumes the type of dynamic process and observation process but leaves the parameters to be determined, which is differ from the case discussed above.

The key of Bayesian inference is the Bayes' theorem, which reverses generating and observation process and obtains estimation process:

$$p(w|y) = \frac{p(y|w)(w)}{p(y)},$$

where $p(w|y)$ is a posterior distribution, $p(y|w)$ corresponds to observation process, $p(w)$ corresponds to a prior knowledge and $p(y)$ is a normalized constant. The Bayesian inference can easily "deepen" the modeling as

$$p(w, \theta|y) \propto p(y|w)p(w|\theta)p(\theta).$$

Here, we can estimate the observation process and the latent dynamical structure through the posterior distribution.

The main problem of Bayesian framework is to calculate the posterior distribution. One of the difficulty arise from the integration over the parameter in normalized constant:

$$p(y) = \int p(y|w)p(w)\mathrm{d}w.$$

One can see that the integration is generally intractable. Here, many techniques to solve the problem are invented: approximation method such as the Laplace approximation and the variational Bayesian frameworks, and sampling method such as Markov chain Monte Carlo method and its extensions.

We call the Bayesian inference as frameworks. Unlike the Kalman filter, the Bayesian inference is not an algorithm but methodology, and one of the important

part to handle this is to model the phenomena behind the data. Once we can model the joint probability, we can apply one of the existing method to it. Clearly, a performance of the Bayesian inference depends on how accurate the probabilistic model is. Therefore, deep insight into the target phenomena is required.

1.2.3 Sparse modeling

When one scarcely know the dynamical and observation process, sparse modeling might be helpful. It assumes that the diversity of the data arises from combination of a few, or sparse, latent elements. For observation $\boldsymbol{y} \in \mathbb{R}^N$, latent variables $\boldsymbol{x} \in \mathbb{R}^M$ and bases $\boldsymbol{W} \in \mathbb{R}^{N \times M}$, we assumes that the observation process is as follows,

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \text{noise}.$$

When N is sufficiently larger than M, we can solve the equation by applying least squares method or its variant. As is often the case, however, when the N is smaller than M, the above equation becomes ill-posed and can not be solved unless any constraint is imposed.

The sparsity is a key concept to find the solution [6]. When $\boldsymbol{x}$ contains many zero elements, in other words, non-zero element in $\boldsymbol{x}$ is sparse, we can resolve the equation. This fact brings us some idea that if we prepare large enough bases to represent the observation, the sparse modeling may detect relevant bases to the data. Obviously, the choice of bases directly affects the performance of the sparse modeling. Dictionary learning is a method to learn the basis only from the data set. One of the most impressive study about the sparsity and the dictionary learning was done by Olshausen and Field [7]. They found the bases which the natural image could be sparsely reconstructed with and which were alike Gabor-filter. The Gabor-filter resembles receptive field of simple cells in the visual cortex, which suggest that brain processes stimulation in a sparse manner. Furthermore, the filter are adopted for image compressing techniques [8].

Learning basis for the sparse representation can be written as matrix decomposition of the form,

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{X},$$

where, $\boldsymbol{Y}$ is the observed data set, $\boldsymbol{W}$ is the bases, and $\boldsymbol{X}$ is the sparse representation of $\boldsymbol{Y}$ with $\boldsymbol{W}$. Note that, since the variables in the right hand side of the equation are both undetermined, some criteria for decomposing are required. Typical examples are the sparse dictionary learning [9], low-rank approximation such as truncated singular value decomposition, non-negative matrix factorization [10, 11], and so on.

For the time-series data, not only the sparseness but also dynamical structure is required. The singular value decomposition, for instance, does not change its

essential properties such as singular values, left and right singular vector, with exchanging between the columns or rows of the data matrix. Time-series data, however, the order of data has critical structure, and hence, exchange between the data causes a significant change in the latent structure behind the data. Recently, dimension reduction techniques introduced time-series structure are proposed: multivariate autoregressive (MAR) model with structured prior [12], matrix factorization method using MAR structure, and so on. One of those methods, called dynamic mode decomposition (DMD) [13, 14], rapidly becomes to attract the attention as it can be applied to nonlinear dynamics. This method has a connection with Koopman operator theory [15, 16] in the area of dynamical systems. The operator $\mathcal{K}$ acts on the observation "function" as

$$\begin{aligned}[\mathcal{K}f](x_t) &\equiv (f \circ g)(x_t) \\ &= f(g(x_t)) \\ &= f(x_{t+1}).\end{aligned}$$

Thus, it can be said that the operator make latent variables one step forward and observe. The operator has particular observation such that

$$[\mathcal{K}\varphi](x_t) = \lambda\varphi(x_t),$$

where, φ and λ is a pair of eigenfunction and eigenvalue. Note that, we can transform finite and nonlinear dynamics to (possibly) infinite and linear dynamics with the Koopman operator. In the vector-valued observation case, we can write the dynamics as

$$\begin{aligned}\boldsymbol{f}(x_t) &= \sum_i \boldsymbol{v}_i \varphi_i(x_t), \\ [\mathcal{K}\boldsymbol{f}](x_t) &= \sum_i \boldsymbol{v}_i \lambda_i \varphi_i(x_t),\end{aligned}$$

where, $\boldsymbol{v}_i$ denote the i-th Koopman mode. DMD is said to be able to approximate the Koopman eigenvalue and mode, and to find the dynamical structure only through the data and without any models.

1.2.4 Our aim

In the last of this section, we briefly summarize related works and address what we aim at. As discussed above, there are many frameworks to analyze time-series data. Table 1.1 shows a pair of what situation the model is and what strategy may works well. In this thesis, we mainly tackle the two situation at the bottom of Table 1.1.

Situation	Strategy
Model is known (including the parameters)	Kalman filter and its variants
Model is known (except the parameters)	(Non)linear fitting, Bayesian inference
Model is not known	Sparse modeling; data-driven method

Table 1.1: Related works

1.3 Summary of contributions

The works studied in this thesis have been already published as journal papers or are planned to be submitted to journals. All of the works studies about the time-series data. Here, we review them and discuss our main contributions.

- **Principal component analysis of the associative memory model.** We analyzed the associative memory model with synaptic depression through the principal component analysis. The model discusses the memory retrieval process. Some of the important properties such as memory capacity and basin of attraction, have been analyzed by using the statistical mechanical method. However, analyzing spurious state corresponding to the non-retrieval state was known to be difficult to solve. Moreover, memory state and spurious state were seemed to be identical in the point of dynamical property view, although their properties of information process differ from each other. By applying PCA, we showed that the synaptic depression, which has been experimentally found mechanism of synapse, did not affect the memory state but mainly destabilized spurious state and induced the oscillations(Chapter 2).

  Original content can be found in Ref. [17]

- **Dynamic mode decomposition to analyze the neural networks.** From the PCA study, the limit cycle formation in the spurious state was suggested. PCA, however, does not take the time structure into account, although it found clear structure. In this work, we focused on the dynamic mode decomposition (DMD), which is originally developed in the field of fluid dynamics. It decomposes the data into the spatio-temporal modes. Moreover, dimensionality reduction can be carried out by applying the sparse-modeling based method, sparsity-promoting dynamic mode decomposition. This works found the clear difference among the memory state, spurious state and oscillations in spurious state. Each neurons have the "phase" in the case of oscillation. This phase and synaptic depression possibly cause the oscillation in the model (Chapter 3).

  This work is planned to submit to a journal.

- **Dynamic mode decomposition for the coherent phonon signals.** We now recognize that DMD is a efficient tools to analyze time-series data. Ultrashort laser pulse can excite the lattice vibration in a substance. Coherent phonon (CP) signals contain the substance-specific normal mode oscillations. The initial phase of the oscillations is important to reveal a dynamics of a photoinduced structural phase transition. Conventionally, the Fourier transformation and the Wavelet transformation are applied to the CP signals. These method utilize the quasi-continuous bases in the frequency space. Thus, the uncertainty width appears and it is difficult to detect the initial phase of the modes. On the other hand, DMD can decompose the signal into a few damping oscillation modes. Moreover, the experimental artificial noise (background noise) is usually superposed to the signal. DMD can decompose such noise from the signal. Note that, this work suggests that DMD can be powerful method for signal processing when the signal consists of (damping) oscillations and background noises (Chapter 4).

  This work is also planned to submit to a journal.

- **Bayesian spectral deconvolution for the time-series data.** Spectral data is observed in the wide range of natural sciences, especially, in the case using spectroscopic measurement. The spectral data has multi-peak structure, and thus, it can be decomposed into the sum of single-peak basis functions. The parameters of basis functions are reflected the nature of subject. Bayesian spectral deconvolution is a method based on the Bayesian inference. It can also estimate the number of basis functions. Time-series spectral data reflects the dynamics of the subject. Thus, we extend the Bayesian spectral deconvolution to the form including dynamical structure. Our proposal method is reduced into the original method if the "dynamics " is like a random noise. Thanks to the introducing dynamical structure, we can estimate the parameters more accurately (Chapter 5).

  Original content is found in Ref. [18]

1.4 Overview of this thesis

This thesis is constructed as follows (as shown in Fig. 1.4). First, Chapters 2 and 3 study about dynamics of a neural network model through dimensionality reduction techniques. Effect of the experimental phenomena on the information process is revealed. Second, Experimental data is analyzed in Chapter 4 with the mode decomposition method appeared in Chapter 3. Third, Chapter 5 proposes the frameworks for analyzing time-series spectral data based on the Bayesian method.

Finally, we summarize our thesis, discuss about the works, and describe the future tasks in Chapter 6.

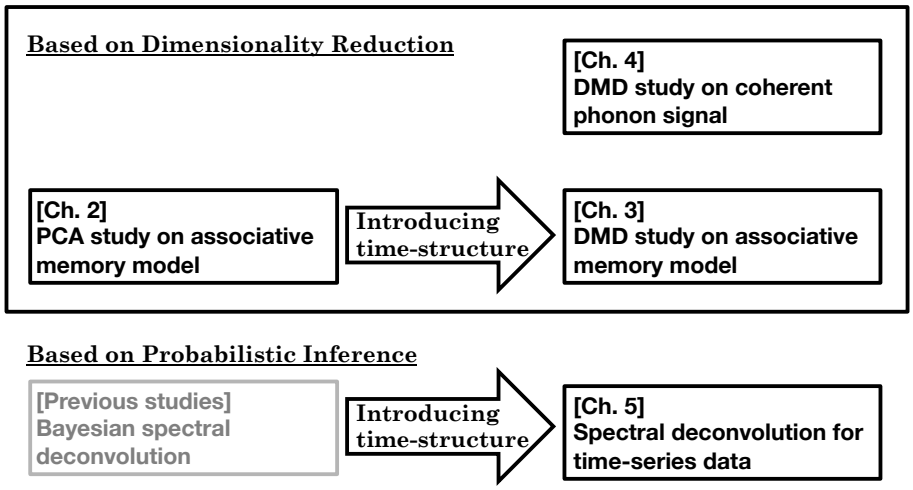


Figure 1.4: Structure of this thesis.

Chapter 2

Principal Component Analysis on the Associative Memory Model

In this chapter and the following chapter, we discuss data-driven approaches to the dynamics of a neural network model. A neural network consists of many simple elements interacting each other, and it can perform information processing through the interaction. Since Hopfield had drawn an analogy between neurons and Ising spin systems [19], statistical mechanical method became to be applied to analyzing properties of neural networks [20, 21, 22]. For instance, Amit *et al.* revealed the loading ratio of the associative memory model, that is, how many memory patterns a neural network can store [21]. However, not all properties of neural networks can be studied analytically. Dynamics of the network can be a one of the such property.

Although analytical study of neural networks is hard to be conducted, we can simulate the model behavior on computer, which may brings us large amount of neural network data to analyze. Hereby, we think that the data-driven approach is a key technology to study neural networks. In addition, the approach described below may also be useful for experimental data.

2.1 Introduction

We will study the associative memory model, which is a typical neural network model and deals with memory retrieval processes. Applying the Hebbian rule enables us to design "energy landscape" on the neural network state. In other words, the network has discretely distributed fixed-point attractors which correspond to memory patterns [19, 23, 24]. Given an initial state near a memory pattern, the network state generally converges on the memory pattern. However, undesirable meta-stable states would appear when the number of memory patterns is on the

order of the number of neurons, in other words, the extensive loading case. These attractors besides memory pattern are called spurious memory. When one considers the associative memory model as an information processing model, the memory states and spurious states play completely different roles. The memory state and spurious state correspond to the success and failure of memory retrieval, respectively. However, it is difficult to distinguish spurious states from memory states, because their dynamical properties are the same. This indistinguishability is a serious problem in using the associative memory model as an information processing model.

Physiological experiments have shown that the efficacy of synaptic transmission changes in a short period of time. For example, dynamic synapses can decrease their transmission efficacy to postsynaptic neurons according to the activities of presynaptic neurons [25, 26, 27, 28]. This phenomenon is called synaptic depression and is a type of short-term synaptic plasticity. Previous studies have revealed that synaptic depression destabilizes attractors and induces switching phenomena among them when the number of memory patterns is much smaller than the number of neurons, i.e., in finite loading cases [29, 30]. However, the network has many attractors including spurious memories in extensive loading cases. The effects of synaptic depression on these attractors have not been analyzed well. Synaptic depression may induce a difference between memory states and spurious states.

In this study, we apply a machine learning method to simulation data to reveal the effects of synaptic depression on attractors including spurious memories. As a result, we find that synaptic depression does not affect memory states but mainly destabilizes spurious states and induces periodic oscillations. This result suggests that incorporating synaptic depression would improve the performance of the associative memory model as an information processing model. It is also suggested that the data-driven approach can be a useful method to analyze neural networks.

2.2 Model

In this section, we introduce an associative memory model with synaptic depression, consisting of N fully connected McCulloch-Pitts neurons. If the i-th neuron fires at discrete time t, its state is represented by $s_i(t) = 1$; otherwise, $s_i(t) = 0$. The network state at time t is denoted by a variable $\boldsymbol{s}(t) = [s_1(t), s_2(t), \cdots, s_N(t)]$. Each neuron is updated with the following probability:

$$\mathrm{Prob}\left[s_i\left(t+1\right)=1\right] = 1 - \mathrm{Prob}\left[s_i\left(t+1\right)=0\right] = F\left[h_i\left(t\right)\right], \tag{2.1}$$

$$h_i\left(t\right) = \sum_{j\neq i}^{N} J_{ij}\left(t\right) s_j\left(t\right). \tag{2.2}$$

Here, $h_i(t)$ represents an internal potential in the i-th neuron at time t, and $F[h_i(t)]$ is a function taking continuous values on $[0, 1]$. We consider an inverse temperature $\beta = 1/T$ and then we define the function $F[h_i(t)]$ as

$$F[h_i(t)] = \frac{1}{2}\{1 + \tanh[\beta h_i(t)]\}. \tag{2.3}$$

The strength of a synaptic connection between the i-th and j-th neurons is denoted by $J_{ij}(t)$ and given by

$$J_{ij}(t) = \tilde{J}_{ij} x_j(t), \tag{2.4}$$

$$x_j(t+1) = x_j(t) + \frac{1 - x_j(t)}{\tau} - U_{SE} x_j(t) s_j(t), \tag{2.5}$$

where $x_j(t)$ corresponds to the efficacy of synaptic transmission [28, 29] and takes continuous values in the range $0 < x_j(t) \leq 1$. Here, U_{SE} represents the fraction of released neurotransmitter in the absence of depression, and τ is a time constant for the recovery process. The associative memory model with synaptic depression is a type of nonlinear second-order time-delay system due to Eq. (5). The model could oscillate according to parameters.

The absolute strength of synaptic connection, $\tilde{J}_{ij}$, is given by the following Hebbian rule:

$$\tilde{J}_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu. \tag{2.6}$$

Here, p corresponds to the number of memory patterns stored in the network. The ratio of the number of neurons, N, to p is called the loading ratio α, i.e., $\alpha = p/N$. Previous studies theoretically show that synaptic depression induces switching phenomena among memory patterns when p is of constant order with respect to N [29, 30]. However, analyzing the spurious states by a statistical mechanical method is very difficult [24], and thus numerical experiments were performed instead.

The stored memory pattern $\boldsymbol{\xi}^\mu = (\xi_1^\mu, \xi_2^\mu, \cdots, \xi_N^\mu)$ is stochastically provided by

$$\mathrm{Prob}[\xi_i^\mu = \pm 1] = \frac{1}{2}. \tag{2.7}$$

When the state of the i-th element in the μ-th memory pattern is in a firing state, $\xi_i^\mu = 1$; otherwise, $\xi_i^\mu = -1$.

2.3 Results

In this section, we show the results obtained from Monte Carlo simulation with the model defined in Sect. 2. To describe the macroscopic state of a network, we

define an overlap between the μ-th memory pattern $\boldsymbol{\xi}^\mu$ and the network state $\boldsymbol{s}(t)$ as

$$M^\mu(t) = \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \left[2s_i(t) - 1\right]. \tag{2.8}$$

The overlap $M^\mu(t)$ corresponds to the direction cosine between the μ-th memory pattern and the state of the network. First, when the state $\boldsymbol{s}(t)$ completely matches the memory pattern $\boldsymbol{\xi}^\mu$, the overlap $M^\mu(t)$ takes a value of 1. Second, when $\boldsymbol{s}(t)$ is orthogonal to $\boldsymbol{\xi}^\mu$, then $M^\mu(t)$ takes a value of 0. Finally, when $\boldsymbol{s}(t)$ matches the sign inversion memory pattern $-\boldsymbol{\xi}^\mu$, $M^\mu(t)$ takes a value of -1. The initial network state $\boldsymbol{s}(0)$ is given by the probability

$$\text{Prob}\left[s_i(0) = 1\right] = 1 - \text{Prob}\left[s_i(0) = 0\right] = \frac{1 + M_0^\mu \xi_i^\mu}{2}. \tag{2.9}$$

The expectation of the overlap $M^\mu(0)$ provided in this way becomes M_0^μ in the limit of $N \to \infty$. Therefore, M_0^μ corresponds to the initial overlap.

2.3.1 Dependence of overlap on step

In this subsection, we show how synaptic depression affects the dependence of the overlap $M^1(t)$ on step t. Figure 2.1 shows the dependence obtained from Monte Carlo simulation with $N = 5000$, $T = 0.1$, and a loading ratio $\alpha = 0.03$, i.e., $p = 150$. Initial states are provided with the initial overlap M_0^1 by varying the value from 0.1 to 1.0 at intervals of 0.1. The parameters of synaptic depression are $\tau = 40$ and $U_{SE} = 0.0125$.

Figure 1(a) shows the case without synaptic depression from $t = 0$ to 20. When M_0^1 is greater than 0.4, $M^1(t)$ converges to 1 at a large t, meaning that memory retrieval is successful. This final state corresponds to a memory state. On the other hand, when M_0^1 is less than or equal to 0.3, $M^1(t)$ is in equilibrium such that the overlaps $M^1(t)$ do not reach 1. This equilibrium is called a spurious state. Figure 1(c) shows the case with synaptic depression. When M_0^1 is greater than 0.4, the network succeeds in memory retrieval; otherwise, when M_0^1 is less than or equal to 0.3, the network fails in memory retrieval and falls into a spurious state.

From Figs. 1(a) and 1(c), we can see that there is little difference between the dynamics of $M^1(t)$ from $t = 0$ to 20 with and without synaptic depression, which means that synaptic depression does not affect memory retrieval.

Figure 1(b) shows the case over a long time span from $t = 0$ to 1000, in Fig. 1(a). In the memory state, the overlap $M^1(t)$ takes a value of 1.0 and then converges. In spurious state, it takes a value of 0.1 and then converges. Figure 1(d) shows the case with synaptic depression over a long time span from $t = 0$ to 1000. In the memory state, $M^1(t)$ takes a value of 1.0 and converges, the same

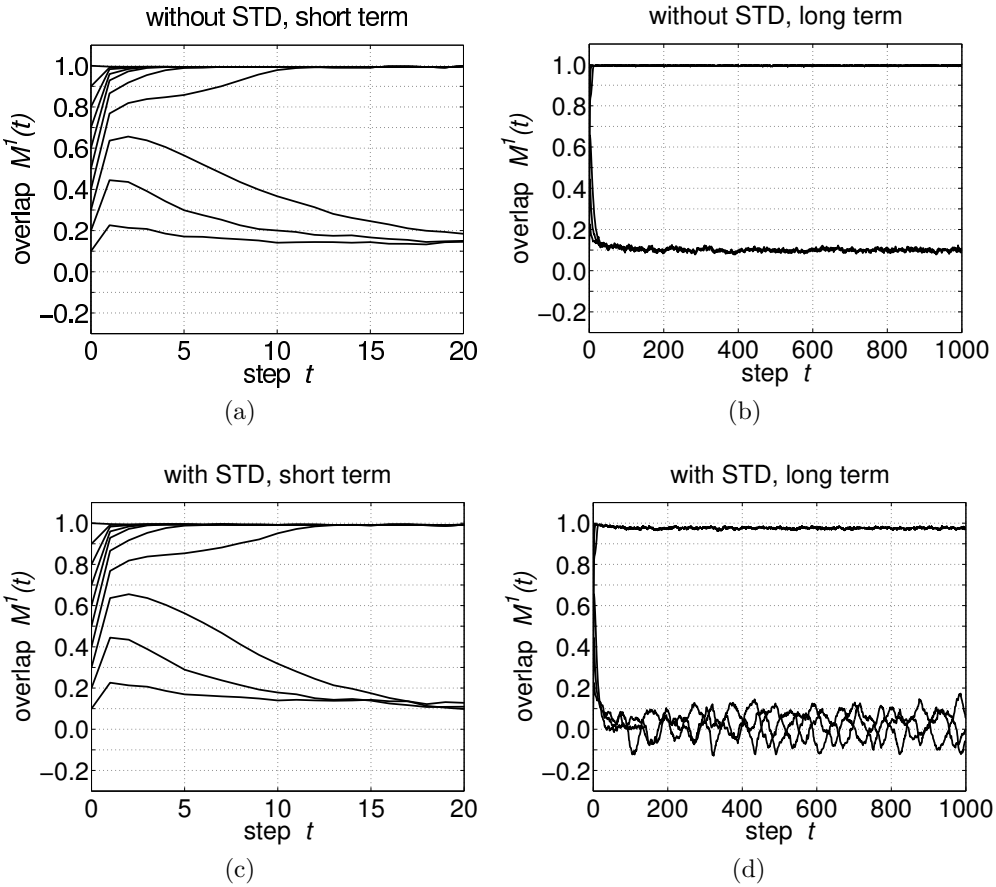


Figure 2.1: Dependence of overlap $M^1(t)$ on step t. The initial overlap M_0^1 takes values of 0.1, 0.2, ..., 1.0. (a) Case without synaptic depression. (b) Conditions in case (a) over a long period. (c) Case with synaptic depression, where $\tau = 40$ and $U_{SE} = 0.0125$. (d) Conditions in case (b) over a long period. Reprinted from Ref. [17].

as the case without synaptic depression shown in Fig. 1(b). On the other hand, $M^1(t)$ does not converge, but oscillates in spurious states.

The results shown in Figs. 1(b) and 1(d) suggest that synaptic depression does not affect memory states, but rather mainly destabilizes spurious states and induces oscillations in overlaps.

2.3.2 Autocorrelation functions and period of network dynamics

The results shown in the previous subsection suggest that synaptic depression does not affect memory states but mainly destabilizes spurious states and induces oscillations. In this subsection, we determine whether or not synaptic depression destabilizes spurious states and induces oscillations. For that purpose, we calculate the autocorrelation function $R(k)$ of the network state $\boldsymbol{s}(t)$. This function is a measure of the dependence of $\boldsymbol{s}(t)$ on previous data and is defined as

$$R(k) = \frac{1}{R_0\,(L-k)} \sum_{t=1}^{L-k} \left[\boldsymbol{s}(t) - \hat{\boldsymbol{s}}\right]^{\mathrm{T}} \left[\boldsymbol{s}(t+k) - \hat{\boldsymbol{s}}\right]. \tag{2.10}$$

Here, L is the number of steps in the network state $\boldsymbol{s}(t)$. Because we can see that $R(k)$ is a type of correlation function between $\boldsymbol{s}(t)$ and k-delayed $\boldsymbol{s}(t+k)$, k represents the time lag between data. The sample mean of $\boldsymbol{s}(t)$, $\hat{\boldsymbol{s}}$, and the normalization constant R_0 are given by

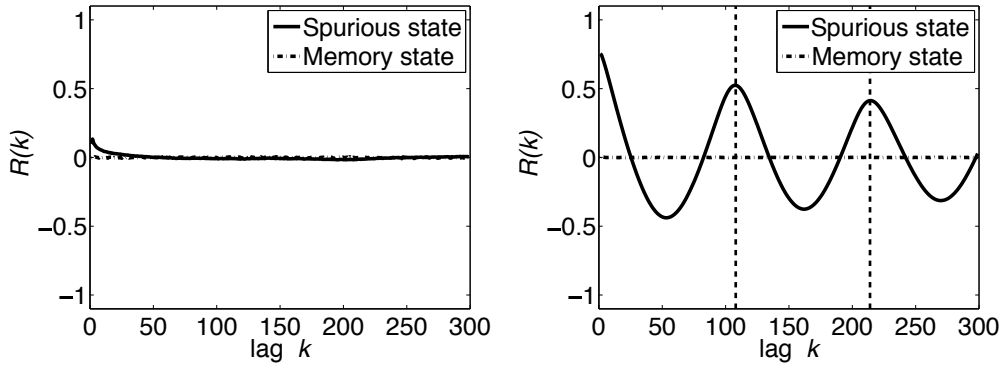
$$\hat{\boldsymbol{s}} = \frac{1}{L} \sum_{t=1}^{L} \boldsymbol{s}(t), \tag{2.11}$$

$$R_0 = \frac{1}{L} \sum_{t=1}^{L} \left[\boldsymbol{s}(t) - \hat{\boldsymbol{s}}\right]^{\mathrm{T}} \left[\boldsymbol{s}(t) - \hat{\boldsymbol{s}}\right]. \tag{2.12}$$

Figure 2.2 shows the dependence of the autocorrelation function $R(k)$ on the lag k. The dashed lines represent $R(k)$ in a memory state, while the solid lines represent $R(k)$ in a spurious state. Note that M_0^1 has values of 1.0 and 0.2 in the cases of a memory state and a spurious state, respectively.

Figure 2.2(a) shows the case without synaptic depression. In both the memory state and the spurious state, $R(k)$ rapidly decays to 0 with increasing k. This means that $\boldsymbol{s}(t)$ is not affected by previous data; rather, its dynamics fluctuates randomly because of the probabilistic update given by Eq. (2.1).

Figure 2.2(b) shows the case with synaptic depression, where $\tau = 40.0$ and $U_{SE} = 0.0125$. In the memory state, $R(k)$ decays to 0 as in the case without synaptic depression. On the other hand, in the spurious state, $R(k)$ has a high positive amplitude at $k = 108$ and 215, meaning that $\boldsymbol{s}(t)$ oscillates with a period



(a) (b)

Figure 2.2: Autocorrelation functions, where the dashed lines correspond to memory states and the solid lines correspond to spurious states. (a) Case without synaptic depression. (b) Case with synaptic depression, where $\tau = 40$ and $U_{SE} = 0.0125$. The orthogonal dotted lines indicate time lags of $k = 108$ and 215. Reprinted from Ref. [17]. © 2014 The Physical Society of Japan.

of approximately 108 in terms of the time lag. From the results shown in Fig. 2, we find that synaptic depression mainly destabilizes spurious states and induces periodic oscillations.

Next, we investigate how the period depends on the strength of synaptic depression. Figure 2.3 shows the dependence of the period on U_{SE}. Here, U_{SE} varies from 0.05 to 0.50 at intervals of 0.05, while τ is fixed at either 5.0, 10, or 15. Each line in the graph represents the mean period obtained from five simulations, and each error bar indicates three standard deviations.

We find that as U_{SE} increases, the period monotonically decreases. We also see that when τ is large, it barely affects the dependence of the period on U_{SE}. Finally, we find that the period decreases as the strength of the synaptic depression increases.

2.3.3 Effects on basin of attraction

It turns out that the success of memory retrieval depends on the initial overlap M_0^1 from Figs. 2.1(a) and 2.1(c). There exists a minimum initial overlap M_0^1, such that the network succeeds in memory retrieval, and this minimum value is called the critical overlap M_C. The region where $M_0^1 > M_C$ is called the basin of attraction, since the network converges on the memory pattern. In this subsection, we investigate the effects of synaptic depression on the basin of attraction. Figure 2.4 shows basins of attraction obtained from 12 simulations. Each cell represents

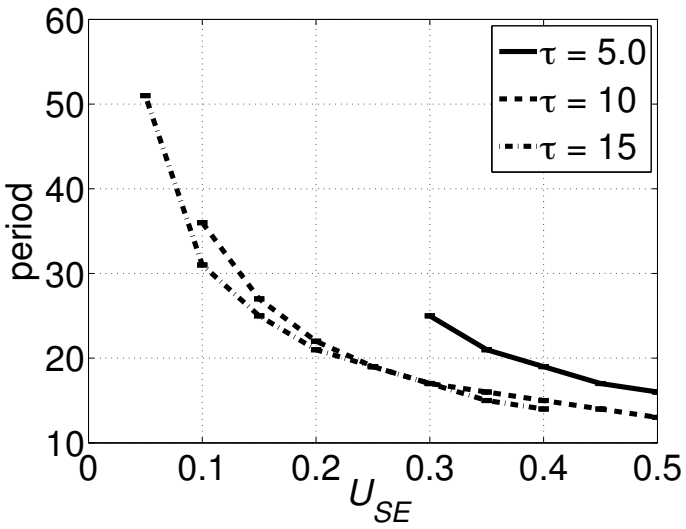
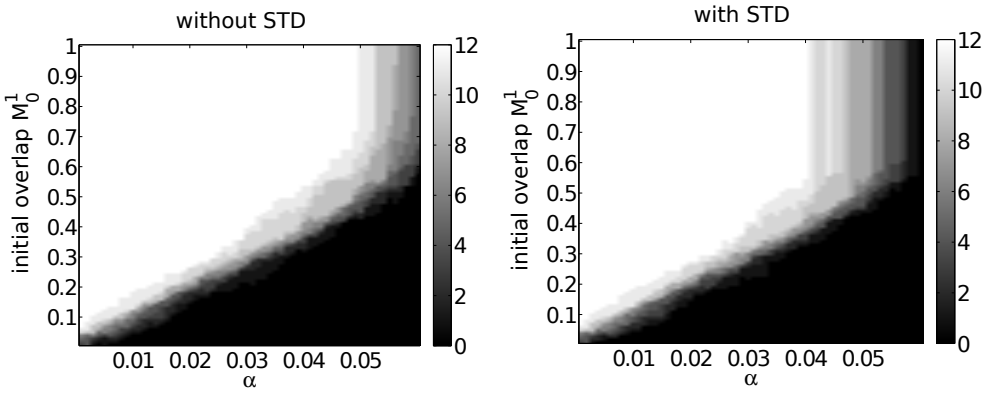


Figure 2.3: Dependence of period on U_{SE} with $\tau = 5.0, 10$, and 15. Each line represents the mean period obtained from five simulations, and each error bar denotes three standard deviations. Reprinted from Ref. [17]. © 2014 The Physical Society of Japan.

the number of successful memory retrievals. In each trial run, α varies from 0.001 to 0.06 at intervals of 0.001, and M_0^1 varies from 0.01 to 1.0 at intervals of 0.01. We define the condition that $M^1(t)$ is greater than or equal to 0.8 at $t = 50$ as a successful memory retrieval. The boundary between the white and black regions corresponds to the critical overlap M_C.

Figure 2.4(a) shows the case without synaptic depression. It turns out that the critical overlap M_C increases with α. Furthermore, M_C markedly increases to 1 at approximately $\alpha = 0.06$. This α is called the memory storage capacity [31, 32, 33]. When α is larger than the memory storage capacity, the network cannot retrieve memory patterns even though it starts from a memory pattern. Figure 2.4(b) shows the case with synaptic depression, where $\tau = 40$ and $U_{SE} = 0.0125$. As in the previous case, we see that M_C increases with α, but in this case, M_C markedly increases to 1 at approximately $\alpha = 0.04$. This decrease in memory storage capacity at finite temperature as a result of synaptic depression was revealed previously [33]. Note that when α is less than 0.04, the dependence of M_C on α shows little difference between the cases with and without synaptic depression, as seen by comparing Figs. 2.4(a) and 2.4(b). This means that when α is less than 0.04, synaptic depression barely affects the basin of attraction.



(a) (b)

Figure 2.4: Basins of attraction obtained over 12 simulations with $T = 0.1$ and $N = 5000$. Each cell represents the number of successful memory retrievals. (a) Case without synaptic depression. (b) Case with synaptic depression, where $\tau = 40$ and $U_{SE} = 0.0125$. Reprinted from Ref. [17]. © 2014 The Physical Society of Japan.

2.3.4 Trajectories of network dynamics

We found that synaptic depression destabilizes stable spurious states and induces periodic oscillations by calculating $R(k)$ in Sect. 3.2. Next, we carry out principal component analysis (PCA) to reveal the dimensionality of the subspace in which the network state $\boldsymbol{s}(t)$ mainly oscillates. We can extract new axes along which the variance in the projected space is maximized by applying PCA, which is a widely used technique for dimensionality reduction. Here, the contribution ratio of each principal component is defined as the ratio of the variance along the principal component to the sum of variances of all principal components. We can find the dimensionality of subspace in which the network mainly oscillates by calculating the contribution ratio. To investigate the dynamical properties of the oscillation induced by synaptic depression, we also project $\boldsymbol{s}(t)$ to the first and second principal components.

We change the initial pattern μ and the initial overlap M_0^μ, and then carry out 750 simulations, i.e., μ varying from 1 to 150 at intervals of 1, and M_0^μ varying from -0.2 to 0.2 at intervals of 0.1. We perform 15000-step simulations, and then count out the first 10000 steps as the initial relaxation. Furthermore, in the case without synaptic depression, the network state in two simulations converges on memory patterns. Thus, we count out these data. We show the result of PCA on the data set obtained from these simulations in Fig. 2.5. Figures 2.5(a) and 2.5(b) show

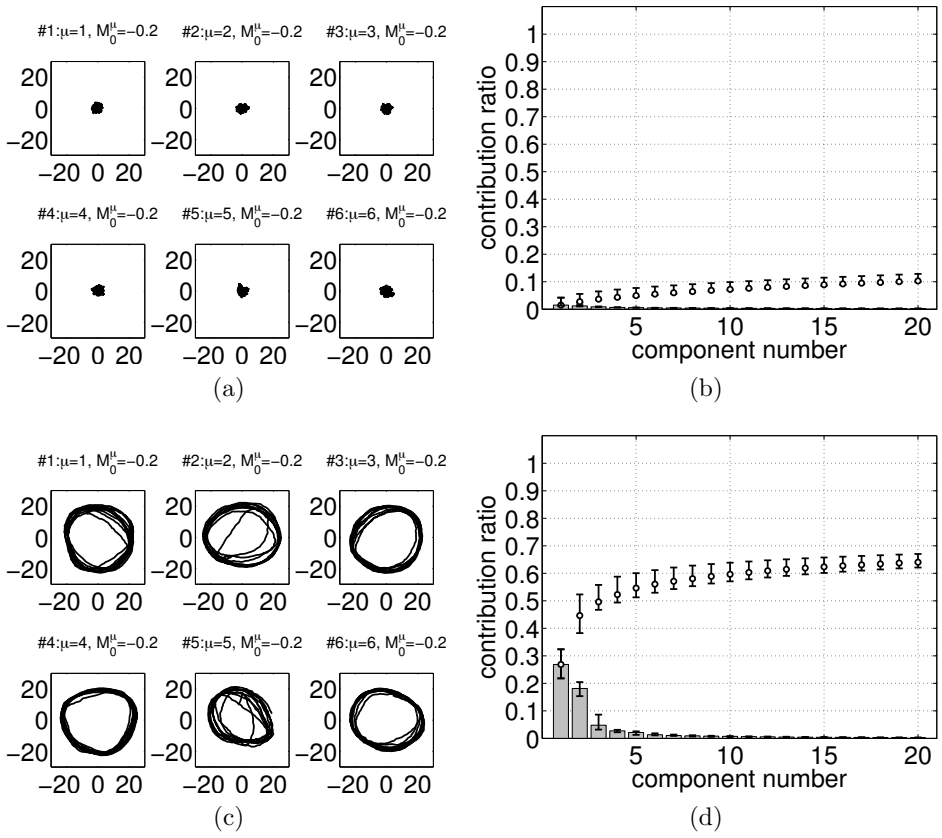


Figure 2.5: Results obtained from PCA on 750 data with different initial conditions, in the cases without (a, b) and with (c, d) synaptic depression, where $\tau = 40.0$ and $U_{SE} = 0.0125$ in the latter case. (a, c) Typical trajectories in a space spanned by the first and second principal components. (b, d) Contribution ratio and cumulative contribution up to the 20th principal component. Gray bars and open circles denote median of the contribution ratio and cumulative contribution, respectively. Error bars represent the 5th–95th percentile range. Reprinted from Ref. [17]. © 2014 The Physical Society of Japan.

the case without synaptic depression. Figure 2.5(a) shows network trajectories on a plane spanned by the first and second principal components. We show the six typical results from 748 trials. Bars and open circles in Fig. 2.5(b) represent the median of the contribution ratio and cumulative contribution, respectively. Error bars represent the 5th–95th percentile range. We find that the networks fluctuate randomly in a high-dimensional subspace, because almost all contribution ratios are less than 0.05. Figures 2.5(c) and 2.5(d) show the case with synaptic depression, where $\tau = 40.0$ and $U_{SE} = 0.0125$. Figure 2.5(c) shows the network trajectories on the plane consisting of the first and second principal components. We show the six typical results from 750 trials. Bars and open circles in Fig. 2.5(d) represent the median of the contribution ratio and cumulative contribution, respectively. Error bars represent the 5th–95th percentile range. We find that the network in spurious states predominantly oscillates on the plane spanned by the first and second principal components, because the contribution ratios of the first and second principal components are higher than those of the other principal components.

From the results shown in Fig. 2.5, we find that synaptic depression changes a stable spurious state into a limit cycle in a low-dimensional subspace.

Next, to investigate the relationship among trajectories, we combine all network state data $\boldsymbol{s}(t)$ obtained from 750 different initial conditions, and then carry out PCA on the data. Figures 2.6(a) and 2.6(b) show the result obtained from PCA in the case without synaptic depression. Figure 2.6(a) shows the distribution of spurious states on the plane spanned by the first and second principal components as a histogram. The distribution is shown as a gray scale image, with black and white corresponding to the smallest and largest numbers of spurious states, respectively. Figure 2.6(b) shows the contribution ratio and cumulative contribution. Bars denote the contribution ratio, and open circles denote the cumulative contribution. We can see that the contribution ratio of the first principal component is higher in this combined data case than in a separate data case. From these results, spurious states in the case without synaptic depression are separated on the plane spanned by the first and second principal components. This result is similar to the behavior of the spin-glass state data of the Sherrington-Kirkpatrick model obtained from PCA [34]. Figures 2.6(c) and 2.6(d) show the result obtained from PCA in the case with synaptic depression, where $\tau = 40.0$ and $U_{SE} = 0.0125$. Figure 2.6(c) shows the distribution of spurious states on the plane spanned by the first and second principal components as a histogram. The distribution is shown as a gray scale image, with black and white corresponding to the smallest and largest numbers of spurious states, respectively. Figure 2.6(d) shows the contribution ratio and cumulative contribution. Bars denote the contribution ratio, and open circles denote the cumulative contribution. The contribution ratio is

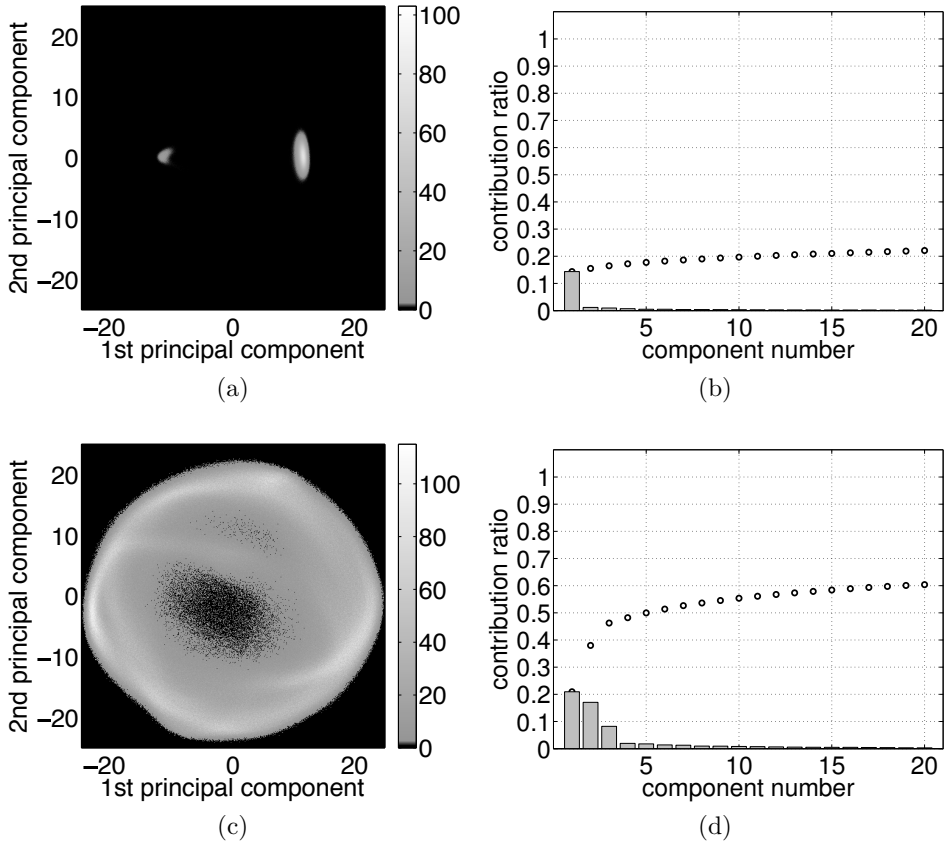


Figure 2.6: Result of PCA on combined data set consisting of 750 different initial conditions, in the cases without (a, b) and with (c, d) synaptic depression, where $\tau = 40.0$ and $U_{SE} = 0.0125$ in the latter case. (a, c) Histogram of plane spanned by the first and second principal components. Brightness represents the number of spurious states. (b, d) Gray bars and open circles correspond to the contribution ratio and cumulative contribution, respectively. Reprinted from Ref. [17]. © 2014 The Physical Society of Japan.

higher up to the third principal component than other principal components. We see that the network in spurious states oscillates in low-dimensional subspace, i.e., approximately three dimensions, in the case with synaptic depression.

2.3.5 Eigenvectors of synaptic weight matrix

From the above discussion, we see that the network in spurious states fluctuates in high-dimensional subspace in the case without synaptic depression, whereas the network oscillates in low-dimensional subspace in the case with synaptic depression. In the former case, the dynamics of memory recall in the associative memory model was discussed qualitatively using eigenvectors of a synaptic weight matrix. Kindo and Kakeya claimed that the linear transformation of a synaptic weight matrix is important for analyzing dynamical properties [35]. Thus, in this subsection, we investigate the relationship between the subspace in which the oscillation induced by synaptic depression occurs and the subspace spanned by eigenvectors of the synaptic weight matrix.

We consider the synaptic weight matrix $\boldsymbol{J}$ whose (i, j) entry is the static weight $\tilde{J}_{ij}$. The eigenvector $\boldsymbol{v}_n$ of $\boldsymbol{J}$ corresponds to the n-th largest eigenvalue λ_n. Here, $\boldsymbol{J}$ has p positive eigenvalues and $(N - p)$ zero eigenvalues. Considering the component of $\boldsymbol{J}$, we see that the subspace spanned by eigenvectors corresponding to positive eigenvalues is composed of memory patterns, and the subspace spanned by eigenvectors corresponding to zero eigenvalues is orthogonal to the subspace spanned by memory patterns.

The network state $\boldsymbol{s}(t)$ is expressed by $\{\boldsymbol{v}_n\}$ as

$$\boldsymbol{s}(t) = \sum_{n=1}^{N} a_n(t)\boldsymbol{v}_n, \tag{2.13}$$

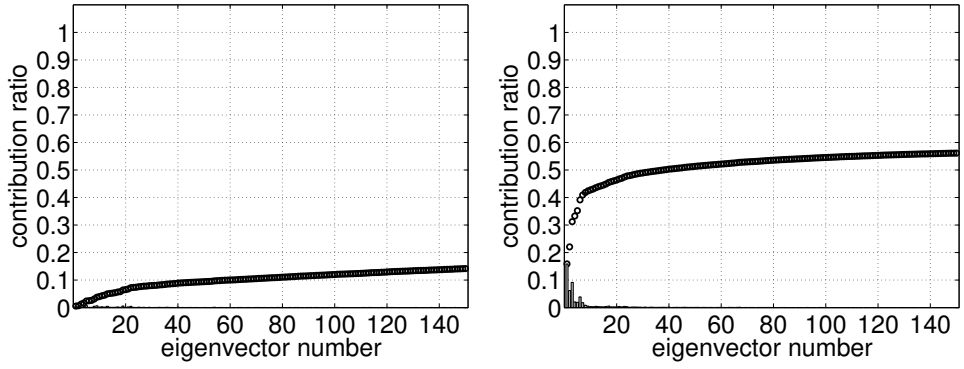
$$a_n(t) = \boldsymbol{v}_n^{\mathrm{T}}\boldsymbol{s}(t). \tag{2.14}$$

The contribution ratio r_n of each $\boldsymbol{v}_n$ is defined as

$$r_n = \frac{\mathrm{var}[a_n(t)]}{\sum_m \mathrm{var}[a_m(t)]}. \tag{2.15}$$

Figure 2.7 shows the contribution ratio r_n and cumulative contribution up to $n = 150$, since $\boldsymbol{J}$ has 150 positive eigenvalues and 4850 zero eigenvalues in this study. Bars and open circles denote the contribution ratio and cumulative contribution, respectively. Figure 2.7(a) shows the case without synaptic depression, and Fig. 2.7(b) shows the case with synaptic depression, where $\tau = 40.0$ and $U_{SE} = 0.0125$. In the case without synaptic depression, we see that spurious states mainly exist in the subspace orthogonal to memory patterns, since the cumulative contribution is approximately 0.14 at $n = 150$. In the case with synaptic depression, however, we

see that spurious states mainly exist in the subspace spanned by memory patterns, since the cumulative contribution is approximately 0.57 at $n = 150$. From the results shown in Fig. 2.7, we see that the oscillation of the network induced by synaptic depression mainly occurs in the subspace spanned by memory patterns.



(a) (b)

Figure 2.7: Contribution ratio and cumulative contribution when the network state is expressed by eigenvector $\boldsymbol{v}_n$ of the static weight matrix $\boldsymbol{J}$. Bars and open circles represent the contribution ratio and the cumulative contribution of $\boldsymbol{v}_n$ up to $n = 150$, respectively. (a) Case without synaptic depression. (b) Case with synaptic depression, when $\tau = 40.0$ and $U_{SE} = 0.0125$. Reprinted from Ref. [17].

2.3.6 Model selection

Here, we apply probabilistic principal component analysis [36]. The network states are assumed to be generated from the following:

$$\boldsymbol{s}_t = \boldsymbol{W}\boldsymbol{z}_t + \boldsymbol{\mu} + \boldsymbol{\epsilon}_t$$

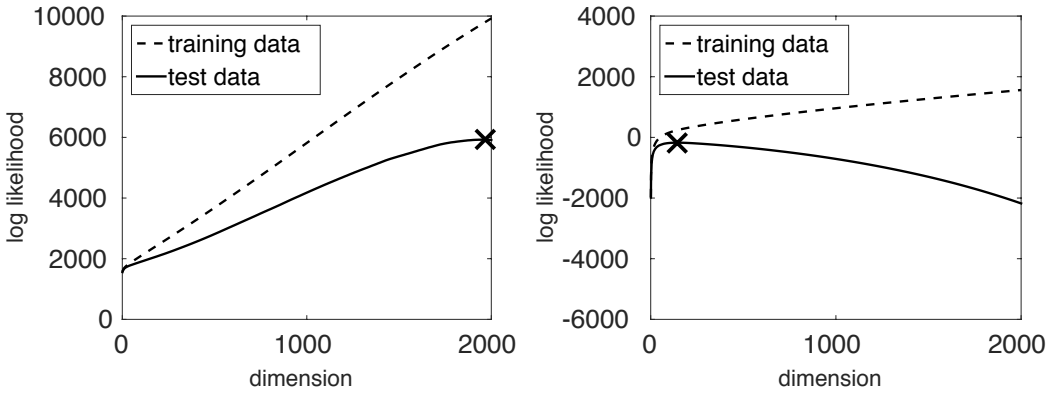
Matrix $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_M)$ denotes basis of the M dimensional subspace, $\boldsymbol{z}_t$ denotes the latent variables, and $\boldsymbol{\mu}$ corresponds to the mean value of observations. Noise term $\boldsymbol{\epsilon}_t$ is assumed to be drawn from $\mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_N)$, and $\boldsymbol{z}_t$ is assumed to be drawn from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_M)$. Then, the log-likelihood of the parameters is determined as

$$\mathcal{L}\Big(\sigma^2, \boldsymbol{\mu}, \boldsymbol{W}\Big) = -\frac{T}{2}\Big\{N \ln(2\pi) + \ln|\boldsymbol{C}| + \mathrm{Tr}\Big(\boldsymbol{C}^{-1}\boldsymbol{S}\Big)\Big\}, \tag{2.16}$$

where $\boldsymbol{C} \equiv \sigma^2 \boldsymbol{I}_N + \boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}$ and $S = \mathrm{cov}[\{\boldsymbol{s}_t\}]$.

In order to find the dimension M maximizing Eq. (2.16), 5 hold cross validation is performed. Figure 2.8 shows mean of the log-likelihood. Solid line and dotted

line correspond to the log-likelihood of test data and training data, respectively. Marks denote the maximum of log-likelihood for the test data. Without synaptic depression, the log-likelihood takes its maximum at $M = 1996$ as shown in Fig. 2.8(a). On the other hand, the log-likelihood takes it maximum at $M = 144$ in the case with synaptic depression as shown in Fig. 2.8(b). This result suggests that synaptic depression reduce the dimension of subspace of spurious state.



(a) The case without synaptic depression. (b) The case with synaptic depression.

Figure 2.8: Dependency of log-likelihood on dimension M.

Now, we discuss the similarity between the subspace selected from PPCA and the subspace spanned by memory patterns.

We consider a matrix consisting of static synaptic weight $vb * J = \left(\tilde{J}_{ij}\right)$. Another matrix $\boldsymbol{V}_M$ contains the first M eigenvectors of $\boldsymbol{J}$ in the descent order of the eigenvalues. Note that, p eigenvectors are related to the memory patterns, and other $(N - p)$ eigenvectors are orthogonal to all the memory patterns. Matrix $\mathbf{V}_M^{\mathrm{T}} \boldsymbol{W}_{\mathrm{ML}}$ denotes the projection from the subspace spanned by $\boldsymbol{W}_{\mathrm{ML}}$ to the subspace spanned by $\boldsymbol{V}_M$. Now, normalized similarity between the subspaces defined as $\det\left(\boldsymbol{V}_M^{\mathrm{T}} \boldsymbol{W}_{\mathrm{ML}}\right)^{1/M}$. Figure 2.9 shows the similarities. The similarity of one thousand pair of subspace generated randomly is also shown. We can see that the similarity between the PPCA subspace and the memory space is significantly high with comparison to those of random subspaces.

2.3.7 Phase diagram

In this subsection, we investigate how the network trajectory depends on the strength of synaptic depression. Figure 2.10(a) shows a phase diagram for this dependence. The horizontal axis corresponds to U_{SE}, and the vertical axis corresponds to τ. U_{SE} varies from 0.05 to 0.5 at intervals of 0.05, and τ varies from

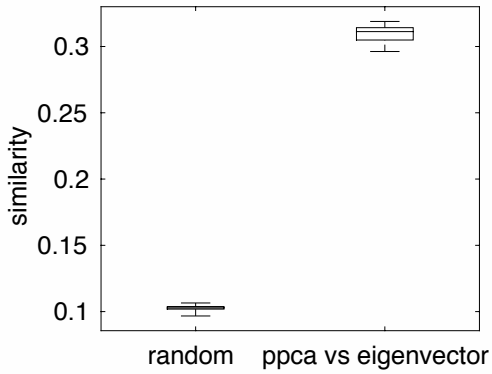


Figure 2.9: Similarity between subspaces.

5.0 to 50 at intervals of 5.0. Each cell represents the network trajectory on the plane spanned by the first and second principal components. The solid and dashed lines correspond to $\gamma = 0.5$ and 9.0, respectively, where γ represents the strength of synaptic depression in equilibrium and is defined as $\gamma = \tau U_{SE}$ [29, 33]. The trajectories have circular motion when γ is more than 0.5 and less than or equal to 9.0.

We now consider the overlap $C(t_m, t_n)$ between the network states $\boldsymbol{s}(t_m)$ and $\boldsymbol{s}(t_n)$, given by

$$C(t_m, t_n) = \frac{1}{N} \sum_{i=1}^{N} \left[2s_i(t_m) - 1\right] \left[2s_i(t_n) - 1\right]. \tag{2.17}$$

When the network state $\boldsymbol{s}(t_m)$ at step t_m completely matches the state $\boldsymbol{s}(t_n)$ at step t_n, the overlap $C(t_m, t_n)$ takes a value of 1. When the network state $\boldsymbol{s}(t_m)$ is orthogonal to $\boldsymbol{s}(t_n)$, $C(t_m, t_n)$ takes a value of 0. Lastly, $C(t_m, t_n)$ takes a value of -1 when the state $\boldsymbol{s}(t_m)$ at step t_m is the reverse of the state $\boldsymbol{s}(t_n)$ at step t_n, i.e., when $\forall i \in \{1, 2, \ldots, N\}, 2s_i(t_m) - 1 = -[2s_i(t_n) - 1]$.

Figures 2.10(b)–2.10(d) show $C(t_m, t_n)$ as heat maps representing the results for different parameter combinations. Note that orthogonal elements take a value of 1 since they denote $C(t_n, t_n)$. Figure 2.10(b) shows the case with $\tau = 5$ and $U_{SE} = 0.05$, where no circular motion is observed. Because the values of $C(t_m, t_n)$ are almost 1, the network converges. Figure 2.10(c) shows the case with $\tau = 20$ and $U_{SE} = 0.1$, where the circular motion is observed. In this case, periodic behavior occurs. Finally, Fig. 2.10(d) shows the case with $\tau = 50$ and $U_{SE} = 0.5$, where no circular motion is observed. Here, the overlaps $C(t_m, t_n)$ take values of approximately 0 except for orthogonal elements, meaning that the network moves randomly.

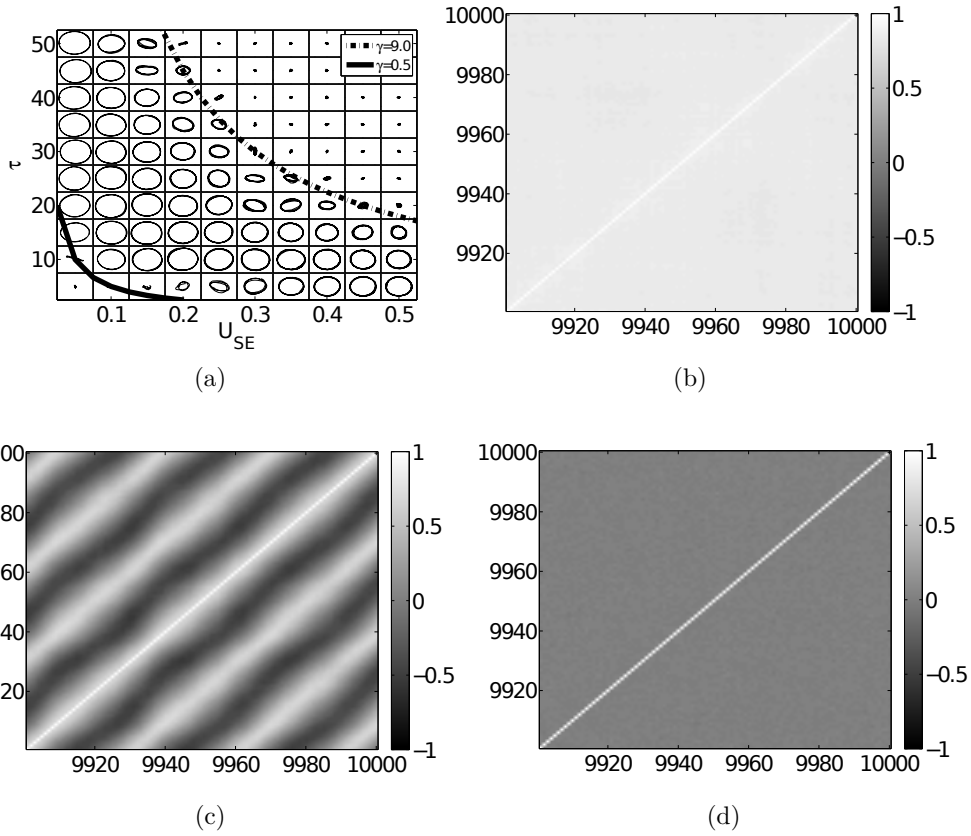


Figure 2.10: (a) Phase diagram for network dynamics in spurious states. (b-d) Overlaps between network states, in the cases of (b) $\tau = 5.0$ and $U_{SE} = 0.05$, (c) $\tau = 20$ and $U_{SE} = 0.1$, and (d) $\tau = 50$ and $U_{SE} = 0.5$. Reprinted from Ref. [17].

Here, we summarize the observed results. First, spurious states in a region where the synaptic depression is weak remain in equilibrium. Second, circular motion begins when the synaptic depression becomes stronger. Third, the network moves randomly in a region of very strong synaptic depression.

2.4 Discussion and Conclusion

In this chapter, we investigated how synaptic depression affects network behavior in memory states and spurious states through the machine learning method. Since both states are in equilibrium, there is little difference between their dynamics. The associative memory model is defective as an information processing model because of this indistinguishability.

We first investigated the dependence of overlaps between the memory pattern and network state in steps, in the cases with and without synaptic depression. As a result, we found that synaptic depression does not affect memory states but destabilizes spurious states and induces periodic oscillations. The oscillation period decreases as the degree of synaptic depression strengthens. These results show that synaptic depression causes the difference in dynamics between memory states and spurious states.

Note that, the overlaps can be calculated only when the memory patters are given. Next, we attempted to analyze the network dynamics by using PCA, which is a widely applied dimensionality reduction method. We found that synaptic depression induces limit cycle like behavior in the spurious state with synaptic depression. Furthermore, we found that the oscillation mainly occurs in the subspace spanned by eigenvectors of a synaptic weight matrix. Through the PCA, we show that the synaptic depression reduces the dimensionality of the network dynamics. In other words, with the help of synaptic depression, the model might search in the subspace it already knows, i.e. subspace spanned by the memory patterns, even though it cannot retrieve the memory patterns.

While PCA can reveal the subspace in which data mainly lies, it does not assume the dynamics in the data. In the next chapter, we will discuss about the dynamics extracted from the data through a machine learning techniques.

Chapter 3

Analyzing Neural Networks through Dynamic Mode Decomposition

The content of this chapter has been temporally unavailable for a future publication. It will be published before March 2022.

Chapter 4

Coherent Phonon Analysis by Using Dynamic Mode Decomposition

The content of this chapter has been temporally unavailable for a future publication. It will be published before March 2022.

Chapter 5

Time-series Spectral Deconvolution

In the previous chapters, we discussed about methods to extract latent dynamical structure based on the sparse modeling frameworks. In this chapter, we will study about a method using Bayesian framework. The framework is based on the probabilistic modeling of the observed data. Hence, it is necessary to model its generating and observing process. When one does not have sufficient prior knowledge to make a model, it might be useful to apply rather dimensionality reduction techniques discussed the previous chapters than the probabilistic methods.

5.1 Introduction

Revealing the properties of an object from observed data is important in natural sciences. Spectral data are a form of observed data used in a wide range of fields. In general, they have a multiple-peak structure. The properties of an object are reflected in the center position, width, and height of each peak. The number of peaks is also an important feature that describes the structure of the data. For instance, the center of a peak corresponds to the binding energy of an object, and the number of peaks corresponds to the number of energy levels in X-ray photoelectron spectroscopy. Therefore, we can reveal the properties of an object by estimating the parameters of peaks and the number of peaks from an observation.

Spectral deconvolution fits spectrum data with a sum of unimodal basis functions, such as Gaussians [37]. On the basis of Bayesian inference, we can estimate the parameters of peaks and their accuracy by applying a probabilistic formulation. The posterior probability distribution is estimated in Bayesian inference. However, the posterior distribution is difficult to calculate in the case of Bayesian spectral deconvolution. Note that the model nature of Bayesian spectral deconvolution pre-

vents us from deriving an effective algorithm for computing the parameters [38] by applying a variational Bayesian method. Thus, sampling methods such as the Markov chain Monte Carlo (MCMC) method are generally applied to calculate the posterior distribution in Bayesian spectral deconvolution. Nagata *et al.* proposed a framework to efficiently estimate the parameters by applying a replica exchange Monte Carlo method [39]. It has also been shown that the number of basis functions can be estimated objectively from the observed data [40, 41].

In this chapter, we propose an extension of Bayesian spectral deconvolution to time-series data [42, 43, 44, 45]. Time-series spectral data have conventionally been analyzed in two steps [46, 47, 39]. First, spectral deconvolution is independently carried out for each set of time-step data. Second, time-series analysis is applied to the results. However, it is not easy to validate the accuracy of the whole estimation in such multistep analysis, although we can validate the accuracy of each step estimation. Thus, we propose an integrative framework to analyze time-series spectral data by introducing a time-series structure into the ordinary probabilistic formulation of spectral deconvolution and by using a replica exchange method. In this paper, we consider a time-series structure called an autoregressive model with the center positions of peaks [48, 49]. By applying our proposed method to synthetic data, we show that we can estimate the parameters more accurately than by the original Bayesian spectral deconvolution.

5.2 Stochastic Formulation

5.2.1 Time-series spectral decomposition

In this research, we propose a hybrid method to treat time-series spectral data as shown in Fig. 5.1(a). Spectral data have multiple peaks. The center, height, and width of each peak reflect the properties of the subject. Conventionally, spectral decomposition [39] is independently applied to spectral data at each time to estimate parameters. In our research, assuming that there is dynamics behind the spectral data, we propose a method to estimate the dynamical structure behind the time-series spectral data.

In this research, we assume that time-series spectral data are generated with the structure shown in Fig. 5.1(b). First, the parameters of the latent dynamics, $\boldsymbol{W}$ and $\boldsymbol{m}$, are generated. Second, the number of peaks K is given and is assumed not to change with time. Third, the parameters of the peaks, $\theta_t = \{a_{k,t}, \mu_{k,t}, \sigma_{k,t}\}_{k=1}^{K}$, are generated at each time in accordance with the dynamics. Finally, spectral data $\boldsymbol{y}_t$ are observed. Bayes' theorem allows us to invert the generation and observation as described above. We apply the Bayesian inference and estimate the parameter set $\Theta = \{\theta_t\}_{t=1}^{T}$, and the latent dynamics parameters $\boldsymbol{W}$ and $\boldsymbol{m}$ from the observed

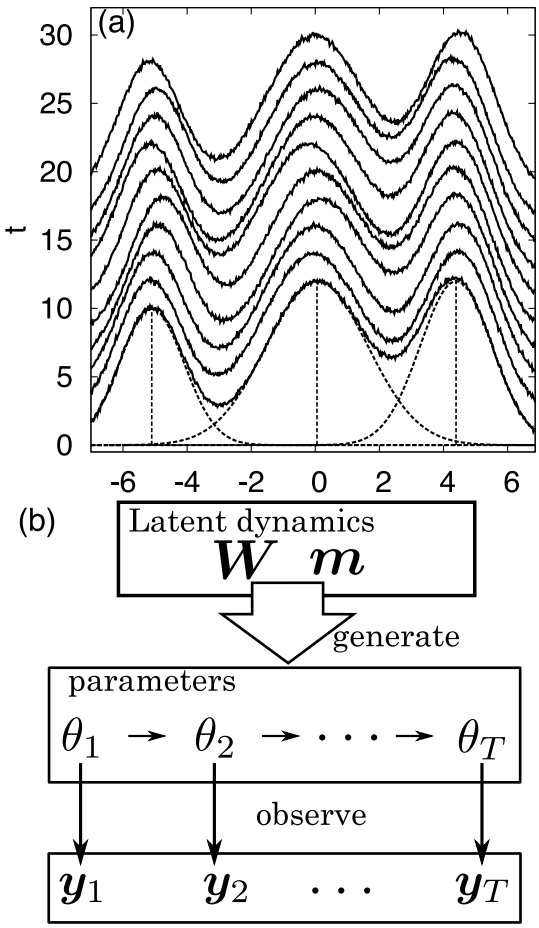


Figure 5.1: Hierarchical structure used in this study. Latent dynamics generates time series of parameters. Spectral data are observed in accordance with the parameters. Reprinted from Ref. [18]. © 2016 The Physical Society of Japan.

spectral data as shown in Fig. 5.1(a).

There have been some previous studies focusing on the movement of spectral peak centers [46, 47]. For the simplest model of such observations, we consider a situation in which the peak centers are time-varying but the peak amplitudes and widths are constant in time. The peak centers $\mu_{k,t}$ are assumed to be generated in accordance with the autoregressive model (AR model):

$$\mu_{k,t} = \sum_{\tau=1}^{d} w_{k,\tau}\mu_{k,t-\tau} + m_k + e_{k,t}. \tag{5.1}$$

Here, $w_{k,1}, \cdots, w_{k,d}$ are coefficients, m_k is constant, and $e_{k,t}$ is a noise drawn from $\mathcal{N}(0, \sigma_{AR}^2)$ [48]. The peak amplitudes $a_{k,t}$ and standard deviations $\sigma_{k,t}$ are time-invariant constants, that is, $a_{k,1} = \cdots = a_{k,T} = a_k$ and $\sigma_{k,1} = \cdots = \sigma_{k,T} = \sigma_k$.

The parameters of the latent dynamics $\boldsymbol{W} = \{w_{k,\tau}\}$ and $\boldsymbol{m} = \{m_k\}$ are independently generated as follows:

$$\begin{aligned} p(\boldsymbol{W}, \boldsymbol{m}) &= p(\boldsymbol{W})p(\boldsymbol{m}) && (5.2)\\ &= \prod_{k=1}^{K} p(m_k) \prod_{k=1}^{K}\prod_{\tau=1}^{d} p(w_{k,\tau}). && (5.3) \end{aligned}$$

From Eq. (5.1), the conditional probability distribution of the k-th peak at time t, $\mu_{k,t}$, given the last d steps of data, $\hat{\boldsymbol{\mu}}_{k,t} = (\mu_{k,t-1}, \cdots, \mu_{k,t-d})^{\mathrm{T}}$, the coefficients of the AR model, $\boldsymbol{w}_k = (w_{k,1} \cdots, w_{k,d})$, and the constant of the AR model, m_k, is

$$p(\mu_{k,t} \mid \hat{\boldsymbol{\mu}}_{k,t}, \boldsymbol{w}_k, m_k) = \mathcal{N}\left(\sum_{\tau=1}^{d} w_{k,\tau}\mu_{k,t-\tau} + m_k, \sigma_{AR}^2\right). \tag{5.4}$$

From the above, the conditional probability distribution of the time series of peak centers, $\{\mu_{k,t}\}$, given the parameters of the AR model, $\boldsymbol{W}$ and $\boldsymbol{m}$, is

$$\begin{aligned} p(\{\mu_{k,t}\} \mid \boldsymbol{W}, \boldsymbol{m}) &= \prod_{k=1}^{K}\prod_{t=1}^{T} p(\mu_{k,t} \mid \hat{\boldsymbol{\mu}}_{k,t}, \boldsymbol{w}_k, m_k) && (5.5)\\ &\propto \exp\left[-\frac{1}{2\sigma_{AR}^2} E_{AR}(\{\mu_{kt}\}; \boldsymbol{W}, \boldsymbol{m})\right] && (5.6)\\ E_{AR}(\{\mu_{kt}\}; \boldsymbol{W}, \boldsymbol{m}) &= \sum_{k=1}^{K}\sum_{t=1}^{T}\left|\mu_{k,t} - \left(\sum_{\tau=1}^{d} w_{k,\tau}\mu_{k,t-\tau} + m_k\right)\right|^2, && (5.7) \end{aligned}$$

where $\mu_{k,0} = \cdots = \mu_{k,1-d} = 0$. We can see that the conditional probability distribution of the parameter set $\Theta = (\theta_1, \cdots, \theta_T)$ given $\boldsymbol{W}$ and $\boldsymbol{m}$ is described as follows:

$$p(\Theta \mid \boldsymbol{W}, \boldsymbol{m}) = p(a_k)p(\sigma_k)p(\{\mu_{k,t}\} \mid \boldsymbol{W}, \boldsymbol{m}). \tag{5.8}$$

Spectral data, $\boldsymbol{y}_t = (y_{1t}, \cdots, y_{Nt})^{\mathrm{T}}$, at each time $t = 1, \cdots, T$, are observed along with the parameters $\theta_t = \{a_k, \mu_{k,t}, \sigma_k\}_{k=1}^{K}$ as shown in Fig. 5.1(b):

$$
\begin{aligned}
y_{it} &= f(x_i; \theta_t) + e_{it}, && (5.9)\\
f(x_i; \theta_t) &= \sum_{k=1}^{K} a_k \phi(x_i; \sigma_k, \mu_{k,t}), && (5.10)\\
\phi(x_i; \sigma_k, \mu_{k,t}) &= \exp\left[-\frac{1}{2\sigma_k^2}\left(x_i - \mu_{k,t}\right)^2\right]. && (5.11)
\end{aligned}
$$

To quantify the difference between the observed data and the fitting curve at each time, the mean squared error (MSE) is defined as

$$
E_t(\theta_t) = \frac{1}{2N}\sum_{i=1}^{N}\left|y_{it} - f\left(x_i; \theta_t\right)\right|^2, (t = 1, \cdots, T), \tag{5.12}
$$

and the MSE averaged over time is defined as

$$
E(\Theta) = \frac{1}{T}\sum_{t=1}^{T} E_t(\theta_t). \tag{5.13}
$$

These definitions are derived when the noise e_{it} in Eq. (5.9) is drawn from $\mathcal{N}(0, \sigma_o^2)$. The conditional probability distribution of y_{it} given θ_t is described as

$$
p(y_{it} \mid \theta_t) = \frac{1}{\sqrt{2\pi\sigma_o^2}} \exp\left[-\frac{1}{2\sigma_o^2}\left|y_{it} - f\left(x_i; \theta_t\right)\right|^2\right]. \tag{5.14}
$$

The conditional probability distribution of $\boldsymbol{y}_t$ given θ_t is described as

$$
\begin{aligned}
p(\boldsymbol{y}_t \mid \theta_t) &= \prod_{i=1}^{N} p(y_{it} \mid \theta_t) && (5.15)\\
&= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_o^2}} \exp\left[-\frac{1}{2\sigma_o^2}\left|y_{it} - f\left(x_i; \theta_t\right)\right|^2\right] && (5.16)\\
&\propto \exp\left[-\frac{N}{\sigma_o^2} E_t\left(\theta_t\right)\right]. && (5.17)
\end{aligned}
$$

Here, the MSE $E_t(\theta_t)$ is derived. When spectral data $\boldsymbol{Y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_T)$ are independently observed, the likelihood of the parameters Θ is described as

$$
\begin{aligned}
p(\boldsymbol{Y} \mid \Theta) &= \prod_{t=1}^{T} p(\boldsymbol{y}_t \mid \theta_t), && (5.18)\\
&\propto \prod_{t=1}^{T} \exp\left[-\frac{N}{\sigma_o^2} E_t\left(\theta_t\right)\right] = \exp\left[-\frac{NT}{\sigma_o^2} E(\Theta)\right]. && (5.19)
\end{aligned}
$$

Here, the time-averaged MSE $E(\Theta)$ is derived. Note that the parameter set Θ minimizing the MSE $E(\Theta)$ matches that maximizing the likelihood $p(\boldsymbol{Y} \mid \Theta)$.

From the above formulation, we can see that the joint probability distribution of the observed spectral data, $\boldsymbol{Y}$, the parameter set, Θ, and the parameters of the latent dynamics, $\boldsymbol{W}$ and $\boldsymbol{m}$, is defined as

$$\begin{aligned} p(\boldsymbol{Y}, \Theta, \boldsymbol{W}, \boldsymbol{m}) &= p(\boldsymbol{Y} \mid \Theta) p(\Theta \mid \boldsymbol{W}, \boldsymbol{m}) p(\boldsymbol{W}) p(\boldsymbol{m}) && (5.20) \\ &\propto \exp\left[-\frac{NT}{\sigma_o^2} E(\Theta)\right] p(\Theta \mid \boldsymbol{W}, \boldsymbol{m}) p(\boldsymbol{W}) p(\boldsymbol{m}). && (5.21) \end{aligned}$$

When spectral deconvolution [39] is independently applied T times, the joint probability distribution is defined as $p(\boldsymbol{Y}, \Theta) = \prod_t p(\boldsymbol{y}_t, \theta_t)$. Note that $\boldsymbol{W}$ and $\boldsymbol{m}$ do not exist, that is, no dynamics is considered.

The posterior distribution of Θ, $\boldsymbol{W}$, and $\boldsymbol{m}$ given spectral data $\boldsymbol{Y}$ is derived as follows by using Bayes' theorem:

$$\begin{aligned} p(\Theta, \boldsymbol{W}, \boldsymbol{m} \mid \boldsymbol{Y}) &= \frac{p(\boldsymbol{Y} \mid \Theta)\, p(\Theta \mid \boldsymbol{W}, \boldsymbol{m})\, p(\boldsymbol{W}) p(\boldsymbol{m})}{p(\boldsymbol{Y})} && (5.22) \\ &\propto \exp\left[-\frac{NT}{\sigma_o^2} E(\Theta)\right] p(\Theta \mid \boldsymbol{W}, \boldsymbol{m})\, p(\boldsymbol{W}) p(\boldsymbol{m}). && (5.23) \end{aligned}$$

We estimate parameters Θ, $\boldsymbol{W}$, and $\boldsymbol{m}$ by calculating the posterior distribution.

5.2.2 Model selection of number of peaks and order of AR model

The number K of Gaussian functions fitting the spectral data and the order d of the AR model are important parameters and are called models. When the model (K, d) changes, the dimensions of parameters $\{\theta_t\}$, $\boldsymbol{W}$, and $\boldsymbol{m}$ change. (K, d) must be objectively determined from the data.

The marginalized posterior of the model (K, d) given data $\boldsymbol{Y}$ is derived as follows:

$$\begin{aligned} p(K, d \mid \boldsymbol{Y}) &= \iiint p(\Theta, \boldsymbol{W}, \boldsymbol{m}, K, d \mid \boldsymbol{Y}) \mathrm{d}\Theta \mathrm{d}\boldsymbol{W} \mathrm{d}\boldsymbol{m} && (5.24) \\ &= \iiint \frac{p(\boldsymbol{Y} \mid \Theta, K) p(\Theta, \boldsymbol{W}, \boldsymbol{m}, K, d)}{p(\boldsymbol{Y})} \mathrm{d}\Theta \mathrm{d}\boldsymbol{W} \mathrm{d}\boldsymbol{m} && (5.25) \\ &\propto p(K, d) \iiint \exp\left[-\frac{NT}{\sigma_o^2} E(\Theta)\right] p(\Theta \mid \boldsymbol{W}, \boldsymbol{m}, K, d) \\ &\quad \times p(\boldsymbol{W} \mid K, d) p(\boldsymbol{m} \mid K) \mathrm{d}\Theta \mathrm{d}\boldsymbol{W} \mathrm{d}\boldsymbol{m}. && (5.26) \end{aligned}$$

The integral in Eq. (5.26) corresponds to the normalization constant of the probability distribution in Eq. (5.23). The free energy is defined as the negative logarithm of the normalization constant,

$$\begin{aligned} F(K,d) &= -\log \iiint \exp\left[-\frac{NT}{\sigma_o^2}E(\Theta)\right] p(\Theta \,|\, \boldsymbol{W}, \boldsymbol{m}, K, d) \\ &\quad \times p(\boldsymbol{W} \mid K, d) p(\boldsymbol{m} \,|\, K) \mathrm{d}\Theta \mathrm{d}\boldsymbol{W} \mathrm{d}\boldsymbol{m}. \end{aligned} \tag{5.27}$$

When the prior distribution $p(K, d)$ of the model is drawn from a uniform distribution, the maximization of the marginalized posterior distribution matches the minimization of the free energy. In our research, the model is selected by minimizing the free energy.

5.2.3 Replica exchange Monte Carlo method

There are two difficulties in carrying out the estimation described above. One is estimating the parameters Θ, $\boldsymbol{W}$, and $\boldsymbol{m}$. Equation (5.22) is difficult to solve analytically. It is also difficult to solve numerically because of its local minima. The other is that the integration in Eq. (5.27) is analytically intractable.

In this research, we solve these two difficulties by using a replica exchange Monte Carlo (REMC) method [50, 51, 39]. This method has two procedures, sampling of the posterior distribution with an MCMC method, such as the Metropolis algorithm, and exchange between different distributions called replicas.

We introduce the inverse temperature β_l to the distribution,

$$\begin{aligned} p\left(\Theta_l, \boldsymbol{W}_l, \boldsymbol{m}_l \mid \boldsymbol{Y}, \beta_l\right) &\propto \exp\left[-\frac{\beta_l NT}{\sigma_o^2} E\left(\Theta_l\right)\right] \\ &\quad \times p\left(\Theta_l \mid \boldsymbol{W}_l, \boldsymbol{m}_l\right) p(\boldsymbol{W}_l) p(\boldsymbol{m}_l). \end{aligned} \tag{5.28}$$

Note that Eq. (5.28) matches the prior distribution with $\beta_l = 0.0$ and the posterior distribution shown in Eq. (5.23) with $\beta_l = 1.0$. We prepare L inverse temperatures and consider the following probability distributions,

$$p\left(\{\Theta_l\}, \{\boldsymbol{W}_l\}, \{\boldsymbol{m}_l\} \mid \boldsymbol{Y}, \{\beta_l\}\right) = \prod_{l=1}^{L} p\left(\Theta_l, \boldsymbol{W}_l, \boldsymbol{m}_l \mid \boldsymbol{Y}, \beta_l\right). \tag{5.29}$$

Then, we attempt to sample the parameters from the above joint probability distribution.

First, we sample parameters drawn from the distribution in Eq. (5.28) by using the Metropolis algorithm at each inverse temperature. The parameters at step k

are denoted as $\Theta_l^{(k)}$, $\boldsymbol{W}_l^{(k)}$, and $\boldsymbol{m}_l^{(k)}$. We choose $\hat{\Theta}_l = \Theta_l^{(k)} + \Delta_\Theta$, $\hat{W}_l = \boldsymbol{W}_l^{(k)} + \Delta_{\boldsymbol{W}}$, and $\hat{\boldsymbol{m}}_l = \boldsymbol{m}_l^{(k)} + \Delta_{\boldsymbol{m}}$. The acceptance ratio r is calculated as follows:

$$r = \min\left\{1, \frac{p\left(\hat{\Theta}_l, \hat{\boldsymbol{W}}_l, \hat{\boldsymbol{m}}_l \mid \boldsymbol{Y}, \beta_l\right)}{p\left(\Theta_l^{(k)}, \boldsymbol{W}_l^{(k)}, \boldsymbol{m}_l^{(k)} \mid \boldsymbol{Y}, \beta_l\right)}\right\}. \tag{5.30}$$

The parameters at step $k+1$ are chosen in accordance with the acceptance ratio r.

Next, the parameters are exchanged between different inverse temperatures β_{l_1} and β_{l_2} after the Metropolis part. In the exchange part, the acceptance ratio is calculated as follows:

$$r_e = \min\left\{1, \exp\left[-\frac{NT}{\sigma_o^2}\left(\beta_{l_1} - \beta_{l_2}\right)\Delta E\right]\right\}, \tag{5.31}$$

$$\Delta E = E\left(\Theta_{l_1}\right) - E\left(\Theta_{l_2}\right). \tag{5.32}$$

The REMC method can avoid the initial-value dependencies of sampling because it allows parameters to be exchanged between different probability distributions.

It also enables us to efficiently calculate the free energy $F(K,d)$ in Eq. (5.27). We consider the integral value f_β,

$$f_\beta \equiv -\log \iiint \exp\left[-\frac{\beta NT}{\sigma_o^2} E\left(\Theta\right)\right] p(\Theta, \boldsymbol{W}, \boldsymbol{m} \mid K, d) \mathrm{d}\Theta \mathrm{d}\boldsymbol{W} \mathrm{d}\boldsymbol{m}. \tag{5.33}$$

Note that $f_{\beta=1}$ matches the free energy $F(K,d)$,

$$f_{\beta=1} = \int_0^1 \frac{\partial f_\beta}{\partial \beta} \mathrm{d}\beta \tag{5.34}$$

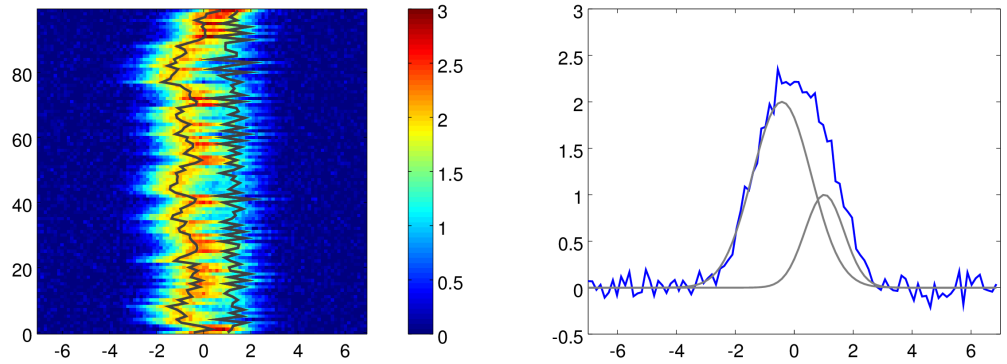
$$= \frac{1}{Z}\int_0^1 \mathrm{d}\beta \iiint \mathrm{d}\Theta \mathrm{d}\boldsymbol{W} \mathrm{d}\boldsymbol{m} \frac{NT}{\sigma_o^2} E(\Theta) \exp\left[-\frac{\beta NT}{\sigma_o^2} E\left(\Theta\right)\right] \times p(\Theta \mid \boldsymbol{W}, \boldsymbol{m}, K, d) p(\boldsymbol{W} \mid K, d) p(\boldsymbol{m} \mid K) \tag{5.35}$$

$$= \int_0^1 \mathrm{d}\beta \left\langle \frac{NT}{\sigma_o^2} E(\Theta) \right\rangle_{p(\Theta, \boldsymbol{W}, \boldsymbol{m} \mid \boldsymbol{Y}, \beta)}, \tag{5.36}$$

where Z is a normalization constant given by

$$Z = \iiint \mathrm{d}\Theta \mathrm{d}\boldsymbol{W} \mathrm{d}\boldsymbol{m} \exp\left[-\frac{\beta NT}{\sigma_o^2} E\left(\Theta\right)\right] \times p(\Theta \mid \boldsymbol{W}, \boldsymbol{m}, K, d) p(\boldsymbol{W} \mid K, d) p(\boldsymbol{m} \mid K). \tag{5.37}$$

Equation (5.35) shows that the free energy $F(K,d)$ can be calculated by integrating the expected value of the MSE $E(\Theta)$ in the distribution given by Eq. (5.28) with



(a) Heat map of spectral data.

(b) Spectral data at $t = 1$.

Figure 5.2: Synthetic data of time-series spectral data.

respect to the inverse temperature β from 0.0 to 1.0. The expected value of $E(\Theta)$ in Eq. (5.28) is obtained from the REMC method, and thus, the free energy $F(K, d)$ can be numerically calculated by using quadrature algorithms with respect to the inverse temperature. As described above, the REMC method estimates the parameters and the selects model.

5.3 Results

In this section, we show the results of applying our proposed method to synthetic data to verify its effectiveness.

5.3.1 Conditions of numerical experiment

We set the true model as $(K, d) = (2, 1)$. One hundred time series $\{\mu_{k,t}\}$ of a peak center are generated by the AR model shown in Eq. (5.1) with the parameters $(w_{1,1}, w_{2,1}) = (-0.6, 0.7)$ and $(m_1, m_2) = (2.0, -0.2)$, and $\sigma_{AR} = 0.3$.

Spectral data $\boldsymbol{y}_t$ are generated from Eq. (5.9) at each time t. The parameters of the Gaussians ϕ_k are $(a_1, a_2) = (1.0, 2.0)$ and $(\sigma_1^{-2}, \sigma_2^{-2}) = (1.5, 1.0)$. The standard deviation of the noise e_{it} is $\sigma_o = 0.1$. The number of data is $N = 100$. The inputs of the fitted function are varied in the range $-7.0 \leq x_i \leq 6.86$ at regular intervals. Synthetic data generated as described above is shown in Fig. 5.2. Figure 5.2(a) shows the heat map of the data. Solid lines corresponds to the movement of the peak centers generated by Eq. (5.1). Figure 5.2(b) shows the spectral data at $t = 1$.

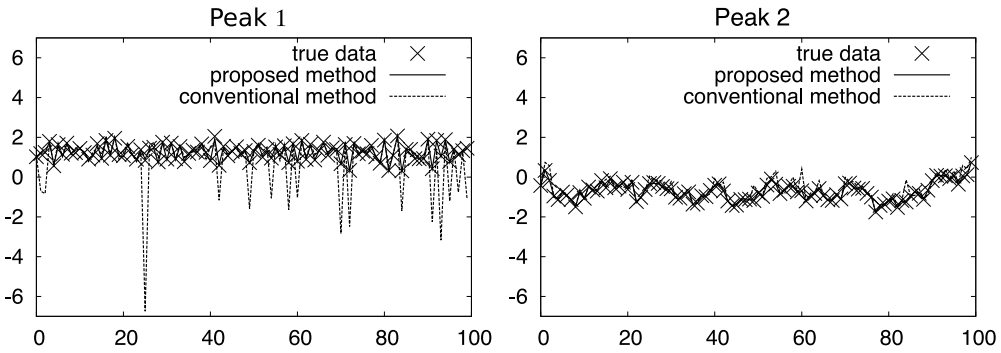


Figure 5.3: Estimation results of peak centers. Solid and dashed lines denote the results obtained from the proposed and conventional methods, respectively. Marks denote true peak centers. Reprinted from Ref. [18]. © 2016 The Physical Society of Japan.

We applied the proposed method to the synthetic data and carried out parameter estimation and model selection. Here, we assumed a continuous uniform prior on the parameters, that is, $p(a_k) \in [0.00, 2.962336]$, $p(\sigma_k^{-2}) \in [0.187403, 18.74028]$, and $p(\mu_k) \in [-7.0, 6.86]$. Note that we sampled σ_k^{-2} instead of σ_k itself, the maximum of a_k corresponds to $\max(\{y_{it}\})$, and the minimum of σ_k^{-2} corresponds to $\{[\max(\{x_i\}) - \min(\{x_i\})]/6\}^{-2}$. We assumed a continuous uniform prior on the AR coefficient $\{w_{k,\tau}\}$ such that the time series of the peak center $\{\mu_{k,t}\}$ remains stationary, i.e., when $d = 1$, $|w_{k,1}| < 1$ and when $d = 2$, $w_{k,2} + w_{k,1} < 1$, $w_{k,2} - w_{k,1} < 1$, and $|w_{k,2}| < 1$. We assumed a discrete uniform prior on the models, i.e., $K = 1, 2, 3$ and $d = 0, 1, 2$. The number L of inverse temperatures is $L = 360$, and each inverse temperature β is set as $\beta_1 = 0.0$ and $\beta_l = 1.045^{l-360}$ $(l = 2, \cdots, 360)$. Note that $\beta_{360} = 1.0$. It is known that the average exchange ratio in a low-temperature limit becomes constant when the inverse temperatures are a geometric progression [52]. We discarded the first 30000 Monte Carlo steps (MCS) as the burn-in period and used the other 50000 MCS to estimate the parameters and select the model.

5.3.2 Results of numerical experiment

First, we estimated parameters with $(K, d) = (2, 1)$. Figure 5.3 shows the estimation results of the peak centers $\{\hat{\mu}_{k,t}\}$. We compared the proposed method with the conventional method, in which we estimated the peak centers independently. Marks denote true values of $\{\mu_{k,t}\}$. Solid and dashed lines denote values estimated by the proposed and conventional methods, respectively. We can estimate the true time series of $\{\mu_{k,t}\}$ by applying the proposed method. To quantify the difference

	Proposed method	Conventional method
$k = 1$	0.001036	0.763886
$k = 2$	0.000297	0.034816

Table 5.1: Mean squared error between true time series of peak centers $\{\mu_{k,t}\}$ and estimated time series $\{\hat{\mu}_{k,t}\}$. Reprinted from Ref. [18]. © 2016 The Physical Society of Japan.

between the true and estimated peak centers, we define the mean squared error,

$$E_k = \frac{1}{2T} \sum_{t=1}^{T} |\hat{\mu}_{k,t} - \mu_{k,t}|^2. \tag{5.38}$$

Table 5.1 shows E_k obtained from the proposed and conventional methods. From Table 5.1, we can see that the mean squared errors obtained from the proposed method are less than those obtained from the conventional method.

Figure 5.4 shows the marginal posteriors of $\{a_k\}$ and $\{\sigma_k^{-2}\}$. The horizontal axis corresponds to the value of each parameter and the vertical axis corresponds to the logarithm of the probability. In each figure, solid, bold dashed, and dashed lines denote marginal posterior distributions, true values, and maximum a posteriori (MAP) estimators, respectively. We can see that each marginal posterior distribution has a sharp peak around a true value and that the MAP estimator also matches the true value. From the above results, the proposed method can accurately estimate the parameters of the amplitude and width of a Gaussian.

Figure 5.5 shows the marginal posteriors of $\{w_{k,\tau}\}$ and $\{m_k\}$. The horizontal axis corresponds to the value of each parameter and the vertical axis corresponds to the logarithm of the probability. In each figure, solid, bold dashed, and dashed lines denote marginal posterior distributions, true values, and MAP estimators, respectively. Figures 5.5(a) and 5.5(b) show the results for m_1 and m_2. We can see that each marginal posterior distribution has a sharp peak around a true value and that the MAP estimator also matches the true value. The estimation accuracy of $\{m_k\}$ is less than that of $\{a_k\}$ and $\{\sigma_k\}$ since the marginal posterior of $\{m_k\}$ is broader than that of $\{a_k\}$ and $\{\sigma_k\}$. Figures 5.5(c) and 5.5(d) show the results for $w_{1,1}$ and $w_{2,1}$. We can see that the sampled data are around the true value since the posterior distribution is narrower than the prior distribution.

Next, we select model (K, d) from the data $\boldsymbol{Y}$ on the basis of Bayesian inference. Table 5.2 shows the values of the free energy $F(K, d)$. We used the results obtained from the REMC method and numerically integrated Eq. (5.35). Here, the model candidates are the combinations of $K = 1, 2, 3$ and $d = 0, 1, 2$. Since model $(K, d) = (2, 1)$ takes the minimum value of the free energy, we can select the

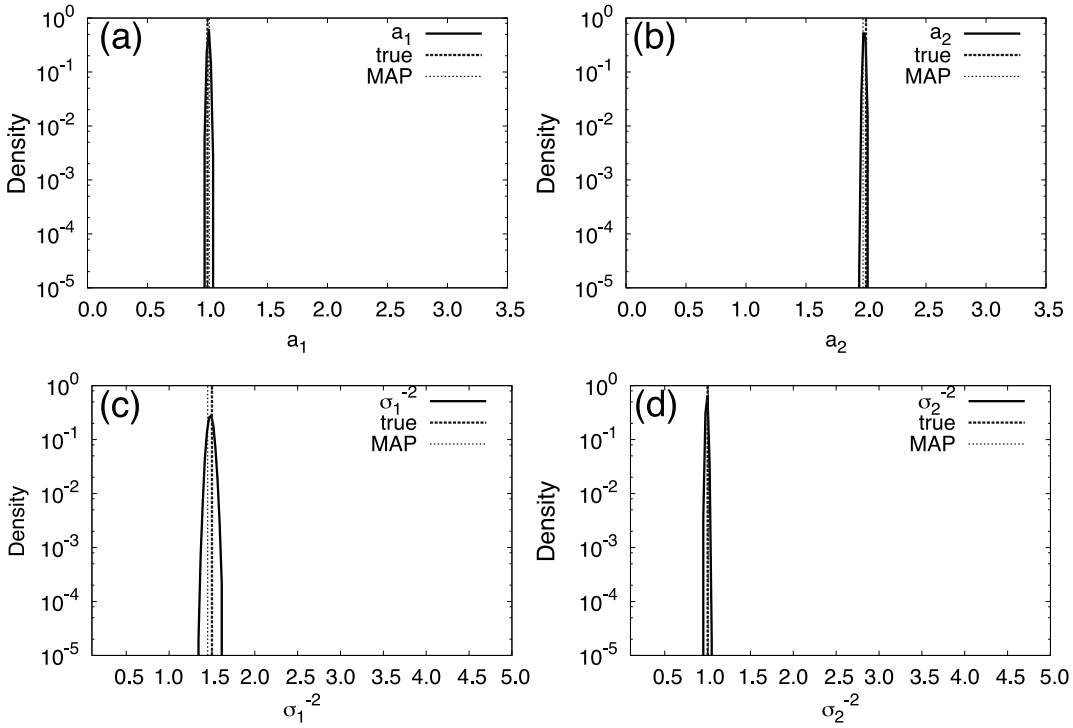


Figure 5.4: Results of parameter estimation. The horizontal axis denotes the value of each variable and the vertical axis denotes the logarithm of the posterior distribution. Solid, bold dashed, and dashed lines denote marginal posterior distributions, true values, and maximum a posteriori (MAP) estimators, respectively. (a)(b) Peak height of each Gaussian. (c)(d) Peak width of each Gaussian. Reprinted from Ref. [18]. © 2016 The Physical Society of Japan.

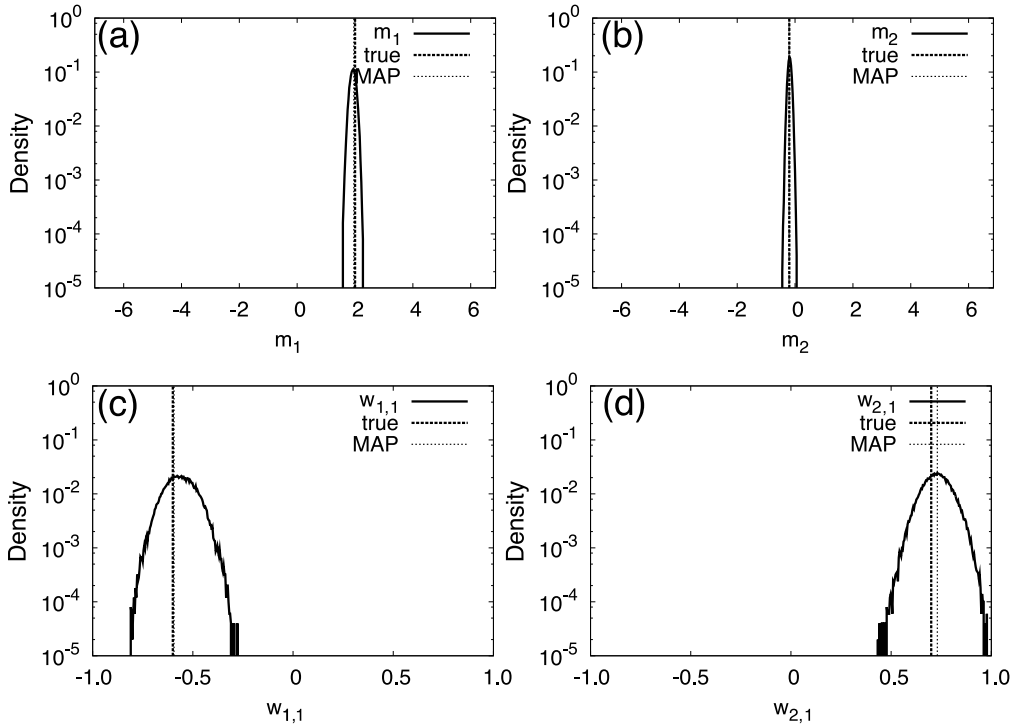


Figure 5.5: Results of parameter estimation. The horizontal axis denotes the value of each variable and the vertical axis denotes the logarithm of the posterior distribution. Solid, bold dashed, and dashed lines denote marginal posterior distributions, true values, and maximum a posteriori (MAP) estimators, respectively. (a)(b) Constant term of AR model. (c)(d) Coefficient of AR model. Reprinted from Ref. [18]. © 2016 The Physical Society of Japan.

	$d=0$	$d=1$	$d=2$
$K=1$	18628.87973	18580.81851	18582.3365
$K=2$	5555.286151	5472.375017	5476.252856
$K=3$	5558.356469	5476.065437	5479.551809

Table 5.2: Dependencies of free energy $F(K,d)$ on the number of peaks K and the order of the AR model d. Reprinted from Ref. [18]. © 2016 The Physical Society of Japan.

	$d=0$	$d=1$	$d=2$
$K=1$	0.0%	0.0%	0.0%
$K=2$	0.0%	95.6%	2.0%
$K=3$	0.0%	2.4%	0.0%

Table 5.3: Dependencies of posterior distribution $p(K,d \mid \boldsymbol{Y})$ on the number of peaks K and the order of the AR model d. Reprinted from Ref. [18]. © 2016 The Physical Society of Japan.

true model correctly through the proposed method. Table 5.3 shows the posterior distribution $p(K,d \mid \boldsymbol{Y})$ calculated on the basis of the free energy. The model is correctly selected.

From the above results, the proposed method can estimate the parameters of the Gaussian and latent dynamics and can select a suitable model.

5.4 Conclusion and Discussion

In this chapter, we proposed a method to analyze time-series spectral data through Bayesian inference. In order to apply the probabilistic method to the observed data, it is necessary to model the generating and observation process based on the prior knowledge. Thus, the framework differs from the dimensionality reduction techniques discussed in the previous chapters, and it is other useful method to analyze the latent dynamical structures.

In the broad range of sciences, it is natural to fit the spectral data by the mixture of basis functions, e.g. Gaussian. Therefore, the estimation problem from the observed data is reduced to the fitting problem. When one consider the observation noise is draw from white Gaussian noise, we can formulate the observation process as a probabilistic model.

Although the previous study proposed the Bayesian estimation method to the

spectral data, the time-series structure were not taken into account. In this chapter, we introduced the dynamical structure derived from the AR process to the generating process of the data, and then, proposed the estimation method for the time-series spectral data.

By using synthetic data, the efficacy of the proposed method was validated. The REMC method was used in order to derive the posterior distribution of the parameters and calculate the free energy to select the model. The proposed method accurately estimated the parameters of the basis functions and the AR process. Moreover, the number of peaks fitting spectral data and the order of the AR model were objectively selected by minimizing the free energy.

In this study, we assumed that only the peak centers have latent dynamics and that other parameters are constant. One possible and important extension of our method is to introduce latent dynamics into other parameters, such as the peak width and peak height.

We suggest several future works. First, we need to introduce a suitable time-series structure into the probabilistic formulation for the time-series data of an object. For instance, Mazet *et al.* introduced a Markov random field model into the peak-center time series as a prior distribution [53]. We can also introduce a time-series structure into the heights or widths of the peaks. Second, the computational cost of the REMC method is high when the number of parameters to be estimated is large. Thus, developing a method with a lower computational cost is another future work. Finally, we need to estimate the levels of AR noise and observation noise. We consider that the estimation can be performed through a hyperparameter estimation framework [54, 55].

Chapter 6

Discussion and Conclusion

In recent years, we have come to be faced with high-dimensional big data. Since the high dimensionality prevents us from understanding the nature of the object, it is important to extract latent (low-dimensional) structure. This is also the case for both static and dynamic data. With comparison to the static data, dynamic data shows the time-correlation among them. This thesis was aimed at extracting the latent dynamics from time-series data, and at helping us to understand the subject.

The associative memory model was studied in Chapters 2 and 3. The model considered the networks consisting of neurons taking the firing or resting state. Memory patterns were corresponded to the firing patterns of the networks and were stored in the network as the strength of connection between the neurons. Recent studies have revealed that the synaptic strength changes in the short period of time according to the presynaptic neuron firing ratio. The steady state of the model with synaptic depression has been analyzed by applying the statistical mechanical method. However, the dynamics is known to be hard to analyze theoretically.

Chapter 2 showed the result of PCA study on the associative memory model with synaptic depression. It was revealed that synaptic depression induced limit cycle attractor in the spurious state which corresponded to the non-retrieval state. Without synaptic depression, the memory and spurious states were indistinguishable since they were both steady state. Formation of the limit cycle only in the spurious state suggested that synaptic depression separate the dynamics of memory state and spurious state from the point of view of information processing.

Dynamic Mode Decomposition (DMD) was applied to the associative memory model in Chapter 3, as the method explicitly took time structure into account. DMD was originally developed in the field of the fluid mechanics. We could decompose the data into the sum of spatio-temporal modes by utilizing DMD. The DMD eigenvalues and the DMD modes were discussed in the chapter. The eigenvalues expressed the time evolution of the mode, and The DMD modes corresponded

the amplitude and the phase of each neurons for each modes. First, we showed that DMD could extract the groups in which the neurons act identically. This result matched the theoretical aspects. Second, the structure of the eigenvalues was found to vary over the memory state, spurious state and the oscillation. This suggested that the DMD was efficient tool to analyze the neural network dynamics. Finally, it was shown that fire patterns were propagated from the neuron to the near neurons when the oscillations occurred.

These result gave us an idea that DMD could be used in the real data which showed oscillating behavior and whose phase information was important to reveal the nature. One suitable data was coherent phonon (CP) signal. The lattice vibration could be induced by ultrashort laser pulse in a substance. Oscillatory changes of the optical constant were observed by using a delayed probe pulse. The CP signal consisted of the finite number of damping oscillations with normal mode frequencies. Initial phases of the damping oscillations were thought to be effective to reveal a dynamics of photoinduced structural phase transition. The CP signal was conventionally analyzed through the Fourier transformation or the Wavelet transformation. These method expanded the signal with quasi-continuous basis in the phase space, such as trigonometric functions and wavelet functions. Hence, the uncertainty width appeared. Moreover, real data contained the experimental artifact noise and fast decaying component near the time-origin. These components deformed the Fourier spectral. Therefore, the estimation of the initial phase became difficult. With comparison to that, we showed that DMD could decompose the signal into a few damping oscillation modes, fast decaying mode and undulation component corresponding to the artificial noise.

Chapters 2 to 4 studied the extraction of latent dynamics based on the dimensionality reduction method. Now, Chapter 5 showed the probabilistic inference based method. One of the advantage of the Bayesian inference was that we could explicitly model the generating and observation process. Thus, the method was vary useful for the measurement sciences. Spectral data was observed over wide range of scientific fields. The width, height and position of the peak were implicitly linked the nature of the object.

Bayesian spectral deconvolution utilized the probabilistic inference framework, it could estimate the parameters with accuracy, and it could determine the number of the basis functions. We expanded the method for the time-series spectral data. The movement of the peak centers were focused on this study. Introducing dynamical structure to the method increased the complexity of the model selection. We also proposed the model selection based on the previous studies. For the synthetic data, the proposed method correctly estimated the parameters and the model.

This thesis focused on the extraction of latent dynamical structure from the

time-series data. We adopted three types of dynamics. First, the associative memory model was picked up as the high-dimensional data, and was analyzed by using the dimensionality reduction methods. Second, CP signals was analyzed in order to tackle the real data, namely, containing the background noise. Finally, as the example of the nonlinear observation, time-series spectral decomposition was proposed.

We think that it is important to refer the applicable range of our study. The equilibrium state and/or the relaxation process are our main issues. Thus, through the whole, we mainly focused on the autoregressive process. Even though it contains the stationary process and random walks, there are other dynamics. One possible and important future work is extension to the transient dynamics among the stationary process. For the nonlinear dynamics, the transient dynamics among the attractors is not uncommon phenomena. Thus, the extension is challenging and useful future task.

Bibliography

[1] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

[2] Zoubin Ghahramani and Geoffrey E Hinton. Variational Learning for Switching State-Space Models. *Neural computation*, 864:831–864, 2000.

[3] Andrew H Jazwinski. *Stochastic Processes and Filtering Theory*, volume 64. Academic Press, 1970.

[4] Simon J Julier and Jeffrey K Uhlmann. New extension of the kalman filter to nonlinear systems. In *AeroSense'97*, pages 182–193. International Society for Optics and Photonics, 1997.

[5] Jun S Liu and Rong Chen. Sequential monte carlo methods for dynamic systems. *Journal of the American statistical association*, 93(443):1032–1044, 1998.

[6] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso Robert Tibshirani. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 58(1):267–288, 1996.

[7] Bruno A Olshausen and DJ Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(13):607–609, 1996.

[8] Adrian S Lewis and G Knowles. Image compression using the 2-d wavelet transform. *IEEE Transactions on image Processing*, 1(2):244–250, 1992.

[9] Kenneth Kreutz-Delgado, Joseph F Murray, Bhaskar D Rao, Kjersti Engan, Te-Won Lee, and Terrence J Sejnowski. Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2):349–396, 2003.

[10] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[11] Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(Nov):1457–1469, 2004.

[12] Will D Penny and SJ Roberts. Bayesian multivariate autoregressive models with structured priors. In *Vision, Image and Signal Processing, IEE Proceedings*, volume 149, pages 33–41, 2002.

[13] Clarence W. Rowley, Igor Mezić, Shervin Bagheri, Philipp Schlatter, and Dan S. Henningson. Spectral analysis of nonlinear flows. *Journal of Fluid Mechanics*, 641:115, 2009.

[14] Peter J. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, jul 2010.

[15] Bernard O Koopman. Hamiltonian systems and transformation in hilbert space. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 17, pages 315–318. National Acad Sciences, 1931.

[16] Marko Budišić, Ryan Mohr, and Igor Mezić. Applied Koopmanism. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 22(4):047510, 2012.

[17] Shin Murata, Yosuke Otsubo, Kenji Nagata, and Masato Okada. Oscillations in Spurious States of the Associative Memory Model with Synaptic Depression. *Journal of the Physical Society of Japan*, 83(124004):124004–1—-124004–8, 2014.

[18] Shin Murata, Kenji Nagata, Makoto Uemura, and Masato Okada. Extraction of Latent Dynamical Structure from Time-Series Spectral Data. *Journal of the Physical Society of Japan*, 85(104003):1–6, 2016.

[19] J. J Hopfield. Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 79, pages 2554–2558, 1982.

[20] Daniel J Amit, Hanoch Gutfreund, and H Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007–1018, 1985.

[21] Daniel J Amit, Hanoch Gutfreund, and H Sompolinsky. Statistical mechanics of neural networks near saturation. *Annals of Physics*, 173:30–67, 1987.

[22] E Gardner. Structure of metastable states in the Hopfield model. *Journal of Physics A: Mathematical and General*, 1047, 1986.

[23] K Nakano. Associatron-A Model of Associative Memory. *Systems, Man and Cybernetics, IEEE Transactions*, 2(3):380–388, 1972.

[24] Masato Okada. Notions of Associative Memory and Sparse Coding. *Neural Networks*, 9(8):1429–1458, 1996.

[25] Alex M Thomson and Jim Deuchars. Temporal and spatial properties of local circuits in neocortex. *Trends in Neurosciences*, 17(3):119–126, 1994.

[26] LF Abbott, JA Varela, K Sen, and SB Nelson. Synaptic depression and cortical gain control. *Science*, 275(5297):221–224, January 1997.

[27] Henry Markram and Misha V Tsodyks. Redistribution of synaptic efficacy between neocortical pyramidal neurons. *Nature*, 382(29):807–810, 1996.

[28] Misha V Tsodyks and Henry Markram. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 94, pages 719–723, 1997.

[29] Lovorka Pantic, JJ Torres, Hilbert J Kappen, and Stan C.A.M. Gielen. Associative memory with dynamic synapses. *Neural Computation*, 2923:2903–2923, 2002.

[30] Yosuke Otsubo, Kenji Nagata, Masafumi Oizumi, and Masato Okada. Instabilities in Associative Memory Model with Synaptic Depression and Switching Phenomena among Attractors. *Journal of the Physical Society of Japan*, 79(8):084002–1–9, 2010.

[31] JJ Torres, Lovorka Pantic, and HJ Kappen. Storage capacity of attractor neural networks with depressing synapses. *Physical Review E*, 66:061910, 2002.

[32] Narihisa Matsumoto, Daisuke Ide, Masataka Watanabe, and Masato Okada. Retrieval property of attractor network with synaptic depression. *Journal of the Physical Society of Japan*, 76(8):084005–1–10, 2007.

[33] Yosuke Otsubo, Kenji Nagata, Masafumi Oizumi, and Masato Okada. Influence of synaptic depression on memory storage capacity. *Journal of the Physical Society of Japan*, 80:084004–1–5, 2011.

[34] Masato Inoue, Koji Hukushima, and Masato Okada. A PCA Approach to Sourlas Code Analysis. *Progress of Theoretical Physics Supplements*, 157(157):246–249, 2005.

[35] Toshiki Kindo and Hideki Kakeya. A Geometrical Analysis of Associative Memory. *Neural Networks*, 11(1):39–51, 1998.

[36] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1997.

[37] Jessica M Sunshine, Carle M Pieters, and Stephen F Pratt. Deconvolution of minerals absorption bands: an improved approach. *Journal of Geophysical Research*, 95(B5):6955–6966, 1990.

[38] Geoffey E Hinton and Drew van Camp. Keeping Neural Networks Simple by Minimizing the Description Length of the Weights. In *Proceedings 6th Ann. Conf. Computational Learning Theory.*, pages 5–13, 1993.

[39] Kenji Nagata, Seiji Sugita, and Masato Okada. Bayesian spectral deconvolution with the exchange Monte Carlo method. *Neural Networks*, 28:82–89, 2012.

[40] Yosihiko Ogata. A Monte Carlo method for an objective Bayesian procedure. *Annals of the Institute of Statistical Mathematics*, 42(3):403–433, 1990.

[41] Mark Jerrum and Alistair Sinclair. Polynomial-Time Approximation Algorithms for the Ising Model. *SIAM Journal on Computing*, 22(5):1087–1116, 1993.

[42] M. Bauer, C. Lei, K. Read, R. Tobey, J. Gland, M. Murnane, and H. Kapteyn. Direct Observation of Surface Chemistry Using Ultrafast Soft-X-Ray Pulses. *Physical Review Letters*, 87:1–4, 2001.

[43] L Nugent-Glandorf, M Scheer, D a Samuels, a M Mulhisen, E R Grant, X Yang, V M Bierbaum, and S R Leone. Ultrafast time-resolved soft x-ray photoelectron spectroscopy of dissociating Br2. *Physical Review Letters*, 87(19):193002, 2001.

[44] José A. Rodriguez. Reduction of CuO in H2: In Situ Time-Resolved XRD Studies. *Catalysis Letters*, 85(3-4):247–254, 2003.

[45] Manami Ogawa, Susumu Yamamoto, Yuka Kousa, Fumitaka Nakamura, Ryu Yukawa, Akiko Fukushima, Ayumi Harasawa, Hiroshi Kondoh, Yoshihito Tanaka, Akito Kakizaki, and Iwao Matsuda. Development of soft x-ray time-resolved photoemission spectroscopy system with a two-dimensional angle-resolved time-of-flight analyzer at SPring-8 BL07LSU. *Review of Scientific Instruments*, 83(023109):1—-7, 2012.

[46] L. Lehr, M.T. Zanni, C. Rischkorn, R. Weinkauf, and D.M. Neumark. Electron Solvation in Finite Systems: Femtosecond Dynamics of IodideâĂć(Water)n Anion Clusters. *Science*, 284(5414):635–638, 1999.

[47] U Korcan Demirok, Gulay Ertas, and Sefik Suzer. Time-Resolved XPS Analysis of the SiO 2 / Si System in the Millisecond Range. *The Journal of Physical Chemistry B*, 108(1):5179–5181, 2004.

[48] Hirotugu Akaike. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, pages 243–247, 1969.

[49] Will D Penny and SJ Roberts. Bayesian Methods for Autoregressive Models. In *Proceedings 2000 IEEE Signal Processing Society Workshop*, volume 1, pages 125–134, Sydney, NSW, 2000.

[50] C J Geyer. Markov chain Monte Carlo maximum likelihood. In *Proceedings23rd Symposium on the Interface*, page 156, 1991.

[51] Koji Hukushima and K Nemoto. Exchange Monte Carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608, 1996.

[52] Kenji Nagata and Sumio Watanabe. Asymptotic behavior of exchange ratio in exchange Monte Carlo method. *Neural Networks*, 21:980–988, 2008.

[53] Vincent Mazet, Sylvain Faisan, Slim Awali, Marc-André Gaveau, and Lionel Poisson. Unsupervised joint decomposition of a spectroscopic signal sequence. *Signal Processing*, 109:193–205, 2015.

[54] David J.C. MacKay. Bayesian Interpolation. *Neural Computation*, 4(3):415–447, 1992.

[55] David J.C. MacKay. Ensemble learning and evidence maximization. In *Proceedings NIPS*, 1995.

Appendix A

Validation for Time-Series Spectral Deconvolution

In Chapter 5, we compare the proposed method and the conventional method for Bayesian spectral deconvolution. As the result shows, the peak center estimated by the conventional method has some irregular values.

In this appendix, we remove those values by hand, and then calculate the mean squared error again. We adopt the criteria $\mu_t < 0$ for the peak 1 in Fig. 5.3. Table A.1 shows the mean squared error for the peak 1. Though the MSE value decrease significantly by removing the irregular value, it is still larger than the MSE of proposed method. Here, we can see that the proposed method is effective to analyze time-series spectral data.

	(Proposed method)	Conventional method
With irregular value	(0.001036)	0.763886
Without irregular value	(0.001036)	0.029022

Table A.1: Mean square error between the estimated and true value of peak center.