博士論文（要約）

Effects of Measurement Noise on Bayesian Spectral Deconvolution: Degenerate or Not

（ベイズ的スペクトル分解に対する測定ノイズの影響：縮退か否か）

徳田　悟

Doctoral Thesis

Effects of Measurement Noise on Bayesian Spectral Deconvolution: Degenerate or Not

Satoru Tokuda

Department of Complexity Science and Engineering, Graduate School of Frontier Sciences, the University of Tokyo

February 2017

Abstract

The quantum nature of spectra allows them to be approximately reduced to the sum of unimodal peaks, whose centers are the energy levels. Observed spectra reflect the *degeneracy*, a case that two or more different states correspond to the same energy, as indicated by solutions of the Schrödinger equation. The degeneracy is removed if the underlying symmetry is broken by some external perturbation. This causes a peak splitting in the observed spectrum. In some delicate cases, such that the gap between the splitting peaks is too small, measurement noise crucially affects the identification of the degeneracy by the observed spectrum. There cannot seem to be several peaks, but only a peak in the observed spectrum, if the magnitude of measurement noise is large, even if Ab initio calculation predicts that there exist the splitting peaks. We state that there surely exist a "degeneracy", caused by measurement, not in the case of degeneracy. In this thesis, we focus on the informational aspects of spectroscopy as *indirect measurements* and clarify the mechanism of "degeneracy". First, we formulate an inverse problem, called *Bayesian spectral deconvolution*, that calculate energy levels from an observed spectrum. We modify Bayesian spectral deconvolution to be applicable even in the case that the noise variance of the observed spectrum is unknown as Bayesian inference originally does so. Second, we take a larger view and focus on the common mathematical structure between statistical estimation and statistical physics. We show that the "degeneracy" is a phase transition in statistical estimation as in statistical physics. Finally, we focus on the measurement by dispersive spectrometers and its essential limit. There is a well-known measurement limit, called the standard quantum limit, which signal and noise cannot be distinguished. There is also a limit of estimation by way of Bayesian spectral deconvolution, which two individual signals cannot be distinguished, i.e. the "degeneracy" or not. We elucidate these limits are uniformly explained as phase transitions in statistical estimation.

Acknowledgments

This thesis is the summary of my works in the Ph. D. course of the Department of Complexity Science and Engineering, Graduate School of Frontier Sciences, the University of Tokyo. Many people have helped me conduct this thesis.

First of all, I would like to thank my supervisor Prof. Masato Okada for his support. My perspective is influenced to no small extent by his unique insight. Actually, I started my research in a different context but miraculously linked up with his conjecture, mentioned in his previous work, as a result. I also would like to thank my mentor Dr. Kenji Nagata for his technical advice. The ideas of my Ph. D. studies are undoubtedly based on my master's ones introduced by him.

My special thanks go as well to the all the members of my Ph. D. committee, Prof. Masato Okada, Prof. Hiroyuki Shinoda, Prof. Takehiko Sasaki, Prof. Koji Hukushima, and Prof. Issei Sato. Their valuable and comprehensive comments helped to improve the draft of this thesis.

I would like to express my gratitude to one of my collaborators Prof. Ichiro Akai. Passionate discussions with him have been illuminating experimentalist's viewpoint, which is the essential piece of my works. I also would like to show my greatest appreciation to Prof. Masayuki Ohzeki and Prof. Chihiro Nakajima. They gave me not only some valuable comments on my works from the viewpoint of statistical mechanics but also courage and chance to spend my life in academia.

I am deeply grateful to all of my lab members and alumni. They have made my research life pleasant and fruitful. Without the existence of my peers Mr. Ryo Karakida and Mr. Shin Murata, I would not have entered the Ph. D. course. Discussions with Mr. Takeshi Ideriha and Mr. Toshiki Nagasaki, who are my junior fellows and my collaborators, have always been providing valuable viewpoints.

Last but not least, I owe a very important debt of gratitude to my family. They have always trusted my choice and encouraged me. I have only one regret about that My grandmother Rei, who is my foster parent, has passed away in the fall before last. Without her, I would not be who I am now and this thesis would not exist. I pray that her soul rests in peace and dedicate this thesis to her.

Contents

Chapter 1

Introduction

1.1 Background

Spectroscopy is at the heart of all sciences concerned with matter and energy, such as astronomy [1–6], catalyst chemistry [7–11], solid-state physics [12–21], structural biology [22–30], surface science [31–33], planetary science [34–37], and plasma physics [38–41]. In each field of science, various spectroscopic methods have been developed and utilized for identifying and quantifying materials, whose component atoms and molecules have unique spectra.

Historically, Newton introduced the word *spectrum* to describe the rainbow of colors that combine to form white light in 1666 almost concurrently with his discovery of the universal gravitation in 1665. Fraunhofer invented the origin of dispersive spectrometers in the early 1800s. Balmer discovered an empirical formula for the visible spectral lines of the hydrogen atom in 1885. His formula and its extensions revealed the discrete nature of energy and contributed to the establishment of quantum mechanics. Now spectroscopic measurements are explained as interactions between light and matter in terms of quantum mechanics and provide an understanding of the physical and chemical properties of the matter.

An electromagnetic spectrum indicates the electronic states and the atomic kinetics. The quantum nature of spectra allows them to be approximately reduced to the sum of unimodal peaks (such as Lorentzian peaks, Gaussian peaks, and their convolutions), whose centers are the energy levels from the semiclassical viewpoint [42]. The peak intensity is proportional to both the population density of the atoms or molecules and their transition probabilities. The Lorentzian peak width indicates the lifetime of the eigenstate due to the time-energy uncertainty relation. The Gaussian peak width indicates the Doppler effect caused by the

atomic kinetics and depends on temperature. These pieces of information about the electronic states or the atomic kinetics are obtained by identifying peaks from spectra.

Observed spectra reflect the *degeneracy*, a case that two or more different states correspond to the same energy, as indicated by solutions of the Schödinger equation. The degeneracy is removed if the underlying symmetry is broken by some external perturbation. This causes a peak splitting in the observed spectrum. There are several types of splits classified by the external perturbation, e.g. the magnetic field (Zeeman effect), the electric field (Stark effect), the crystal field or the ligand field (Jahn-Teller effect). Some structural phase transitions, influenced by pressure or temperature, also cause the splitting. Identifying by the observed spectra whether the energy level is degenerate or not is directly linked to an understanding what properties are there in the system of interest.

In some delicate cases, such that the gap between the splitting peaks is too small, measurement noise crucially affects the identification of the degeneracy by the observed spectrum. There cannot seem to be several peaks, but only a peak in the observed spectrum, if the magnitude of measurement noise is large, even if Ab initio calculation predicts that there exist the splitting peaks. We state that there surely exist a "degeneracy", caused by measurement, not in the case of degeneracy. Spectroscopic measurements are based on the interactions between light and matter and essentially include fluctuations, measurement noise. Ab initio calculation does not especially emulate the measurement because there is the interest not in measurement but in the measured object itself. This is a definitive difference so far between theory and experiment in physics and the related fields.

1.2 Our concepts and preliminaries

In this thesis, we focus on the informational aspects of spectroscopy as *indirect measurements* and clarify the mechanism of the "degeneracy", caused by measurement, in terms of our approach. An indirect measurement is a method to obtain the quantity of interest not directly by measuring itself but by way of measuring the other quantities. This needs to solve the inverse problem of estimating the unmeasurable quantity from the measurable ones based on the relation between them, e.g. calculating the spring stiffness by way of Hooke's law. We simulate the spectroscopic indirect measurement which consists of two steps: (i) obtain the spectrum according to the forward problem with the control parameters cor-

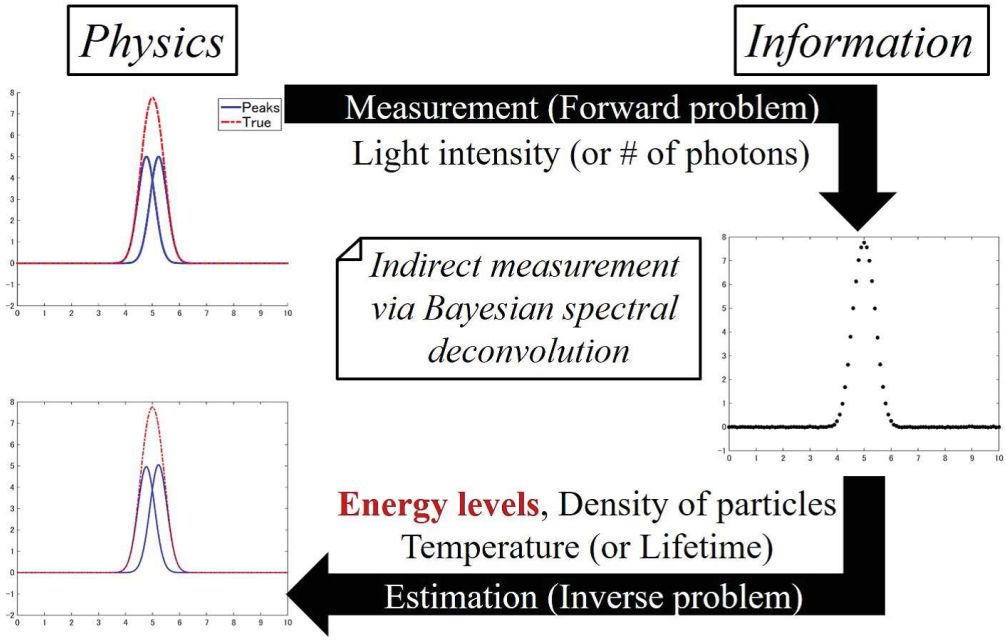

Figure 1.1: Schematic picture of our approach.

responding to the measurement conditions and (ii) solve the inverse problem of estimating the energy levels from the obtained spectrum.

1.2.1 Forward problem

Assume that the forward problem of the spectroscopic (direct) measurement is expressed as

$$y = Y(G(x; w); b), \tag{1.1}$$

where the function $G(x; w)$ of the energy x is the ideal spectrum determined only by the measured system with the physical parameter w. The function $Y(G(x; w); b)$ is the measured spectrum depending both on $G(x; w)$ and on the measurement condition b in general. If Y works as the additive white Gaussian noise, $Y(G(x; w); b)$ is reduced into

$$y := G(x; w) + \varepsilon, \tag{1.2}$$

where ε is the random variable depending on the Gaussian distribution whose mean and variance are respectively 0 and b^{-1}. From the semiclassical viewpoint, the effective model of $G(x; w)$ is expressed as

$$G(x; w) := \sum_{k=1}^{K} a_k \phi_k(x; \mu_k, \rho_k) \tag{1.3}$$

in some cases, where

$$\phi_k(x;\mu_k,\rho_k) := \exp\left[-\frac{\rho_k}{2}\left(x-\mu_k\right)^2\right], \tag{1.4}$$

and $w := \{a_k, \rho_k, \mu_k\}_{k=1}^K$. $\mu_k \in \mathbb{R}$ means the energy level, which is the quantity of interest. $a_k > 0$ and ${\rho_k}^{-1/2} \geq 0$ respectively correspond to the density of particles and temperature in each energy level. If $\mu_k = \mu_{k'} = \mu^*$ and $\rho_k = \rho_{k'} = \rho^*$ for $k \neq k'$, the modes $a_k\phi_k$ and $a_{k'}\phi_{k'}$ are regarded as degenerate. In such cases, $a^*\phi_k(x;\mu^*,\rho^*) = a_k\phi_k(x;\mu_k,\rho_k) + a_{k'}\phi_{k'}(x;\mu_{k'},\rho_{k'})$ with $a^* = a_k + a_{k'}$ holds. Note that our scope is the measurement itself so that we does not consider what physical situation causes the degeneracy and what materials construct the measured system.

1.2.2 Inverse problem

To obtain w from the data set $D := \{X_i, Y_i\}_{i=1}^n$ based on Eqs. (1.2) and (1.4) is one of the nonlinear inverse problems, which are typically ill-posed. Consider the noiseless case for simplicity. If K is known and $n \geq 3K$, then the considered problem is well-posed. Otherwise, this problem is ill-posed. For example, the solution w of $G(x;w) = a_1\phi_1(x;\mu_1,\rho_1)$ are not unique: $\forall w \in W$ given by

$$W := W^{(1)} \cup W^{(2)}, \tag{1.5}$$

$$W^{(1)} := \{w \mid a_k = 0, a_{k'} = a^*, \mu_{k'} = \mu^*, \rho_{k'} = \rho^*\}, \tag{1.6}$$

$$W^{(2)} := \{w \mid a_k + a_{k'} = a^*, \mu_k = \mu_{k'} = \mu^*, \rho_k = \rho_{k'} = \rho^*\} \tag{1.7}$$

are some of the solutions with $K = 2$. The above considerations show that the ill-posedness arises from the variable K. Practically, there naturally exist the measurement noise, which simply makes any problems ill-posed as well. These two factors are the keys to explaining the mechanism of the "degeneracy".

Statistical estimation, especially Bayesian inference, is a better way to solve the ill-posed problem. Bayesian inference is formulated based on Bayes' theorem, expressed as

$$p(B \mid A) = \frac{p(A \mid B)p(B)}{p(A)}, \tag{1.8}$$

where

$$p(A) = \int dB p(A \mid B)p(B) \tag{1.9}$$

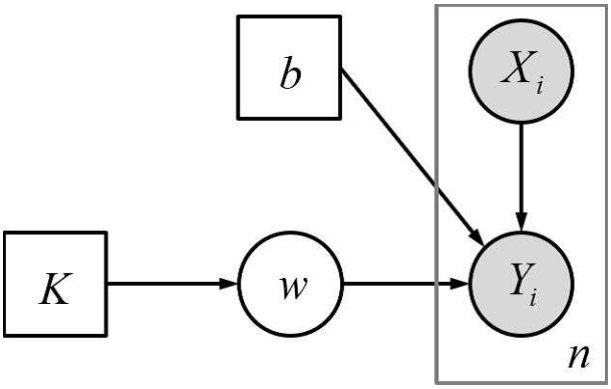

Figure 1.2: Graphical model of Bayesian spectral deconvolution.

for any continuous random variables A and B. $p(B)$, $p(B \mid A)$, $p(A \mid B)$ and $p(A)$ are respectively called the prior distribution, the posterior distribution, the likelihood and the marginal likelihood. In Bayesian inference, the solution w of Eq. (1.2) is derived in the form of the posterior distribution

$$p(w \mid D, K, b) = \frac{p(D \mid w, K, b)p(w, K, b)}{p(D, K, b)} \tag{1.10}$$

$$= \frac{\prod_{i=1}^{n} p(Y_i \mid X_i, w, b)p(w \mid K)}{p(Y^n \mid X^n, K, b)}, \tag{1.11}$$

where $p(y \mid x, w, b) := N(f(x; w), b^{-1})$, $X^n := \{X_i\}_{i=1}^n$, $Y^n := \{Y_i\}_{i=1}^n$ and

$$p(Y^n \mid X^n, K, b) = \int dw \prod_{i=1}^{n} p(Y_i \mid X_i, w, b)p(w \mid K). \tag{1.12}$$

Eq. (1.11) is derived from Eq. (1.10), assuming the dependence between the random variables as sketched in Fig. 1.2. $p(w \mid D, K, b)$ includes all the information on the mean and standard deviation of the element of w, which are depending on and appropriate to the D obtained by the (direct) measurement. This property of Bayesian inference makes the indirect measurement worthy to be called a "measurement". A considerable point is the modeling of $p(w \mid K)$, which means the prior knowledge of w without the observation of D. To express the physical law that w, which is a physical quantity, obeys as $p(w \mid K)$ is most desirable from our point. If such a modeling is difficult, the observers' subjectivity, which drives them to measure w, should be explicitly modeled in the form of $p(w \mid K)$ as a second best way. Whether you consider this way to be natural or not as a scientific method depends on your religion.

In the case that K and b are unknown, they can also be estimated in the form

of the joint posterior distribution

$$p(K, b \mid D) = \frac{p(Y^n \mid X^n, K, b)p(K)p(b)}{p(Y^n \mid X^n)}, \tag{1.13}$$

where

$$p(Y^n \mid X^n) = \sum_K \int dbp(Y^n \mid X^n, K, b)p(K)p(b). \tag{1.14}$$

Assuming that $p(K)$ and $p(b)$ are the uniform distributions defined by any intervals, $p(K, b \mid D)$ is simply proportional to $p(Y^n \mid X^n, K, b)$. In the empirical Bayes approach [43–45], K and b are estimated as the most likely values that maximize $p(Y^n \mid X^n, K, b)$. A similar type of approach, called *Bayesian spectral deconvolution*, was originally proposed by Nagata et al. [46], then applied by Hong et al. [47] and Hagino [48], later extended by Kasai et al. [49] and Murata et al. [50] for the specific situation. However, the formulation of these studies, which b is assumed to be known, does not enable estimating b. We modify Nagata et al.'s framework to be applicable even in the case that b is naturally unknown as Bayesian inference originally does so [51].

The empirical Bayes approach resolves the ill-posedness caused by K and b, which are estimated depending on D. $p(Y^n \mid X^n, K, b)$ naturally embodies Occam's razor: MacKay showed the explicit form for Bayesian linear regression [45]. However, the integration of Eq. (1.12) is generally intractable. Watanabe showed that the asymptotic form of $F_n(K, b) := -\log p(Y^n \mid X^n, K, b)$ is expressed as

$$F_n(K, b) = nL_n(w_0; b) + \lambda \log n + O_p(\log \log n) \tag{1.15}$$

for $n \to \infty$ [52, 53], where

$$L_n(w; b) := -\frac{1}{n} \sum_{i=1}^{n} \log p(Y_i \mid X_i, w, b). \tag{1.16}$$

w_0 is the parameter that minimizes the Kullback–Leibler divergence of $p(y \mid x, w, b)$ from a true distribution, and $\lambda > 0$ is a rational number called the real log canonical threshold (RLCT). The values $L_n(w_0)$ and λ respectively become larger and smaller as K increases. This trade-off works as Occam's razor and ensures that the estimated K is moderate for the given D. In the special case, Eq. (1.15) is reduced into

$$F_n(K, b) = nL_n(\hat{w}; b) + \frac{\dim(w)}{2} \log n + O_p(1), \tag{1.17}$$

known as Bayesian information criterion (BIC), proposed by Schwarz [54]. $\hat{w}$ is the parameter set that minimizes $L_n(w;b)$. Nagata et al.'s framework of Bayesian spectral deconvolution utilizes the above Occam's razor for estimating K in the case that b is known and fixed. They, therefore, do not consider the Occam's razor caused by b.

We explicitly express the Occam's razor caused by b in the form of

$$F_n(K,b) = b\tilde{F}_n(K,b) - \frac{n}{2}(\log b - \log 2\pi), \tag{1.18}$$

where

$$\tilde{F}_n(K,b) = nE_n(w_0) + \frac{\lambda}{b}\log n + \frac{1}{b}O_p(\log\log n), \tag{1.19}$$

$$E_n(w) := \frac{1}{2n}\sum_{i=1}^{n}(Y_i - G(X_i;w))^2. \tag{1.20}$$

Note that we used the relation

$$L_n(w;b) = bE_n(w) - \frac{1}{2}(\log b - \log 2\pi). \tag{1.21}$$

The dependence of $F_n(K,b)$ on K is summarized as $\tilde{F}_n(K,b)$. The values $E_n(w_0)$ and λ respectively become larger and smaller as K increases. This trade-off, whose balancing factor is b, works as Occam's razor. The term $nE_n(w_0)$ dominantly works for large b so that K that minimize $\tilde{F}_n(K,b)$ becomes large, and vice versa. The K appropriate to D is therefore estimated under the b appropriate to D. Such a pair of K and b minimize $F_n(K,b)$, i.e., maximize $p(Y^n \mid X^n, K, b)$.

1.2.3 Effects of measurement noise

We consider the effects of measurement noise on statistical estimation, especially on Bayesian spectral deconvolution. It has been pointed out that Bayesian inference and statistical physics have the common mathematical structure. The generalized formulation of Bayesian inference, which is equivalent to statistical physics' one, was proposed by Watanabe [55]. He also mentioned that there are phase transitions in statistical estimation as in statistical physics. We show that the "degeneracy" is a phase transition in statistical estimation as in statistical physics.

We focus on the measurement by dispersive spectrometers and its essential limit. A dispersive spectrum has the integer-valued intensity, corresponding to the number of photons, with the Poisson noise due to the discrete nature of photon. There is a measurement limit, called the standard quantum limit, which signal and

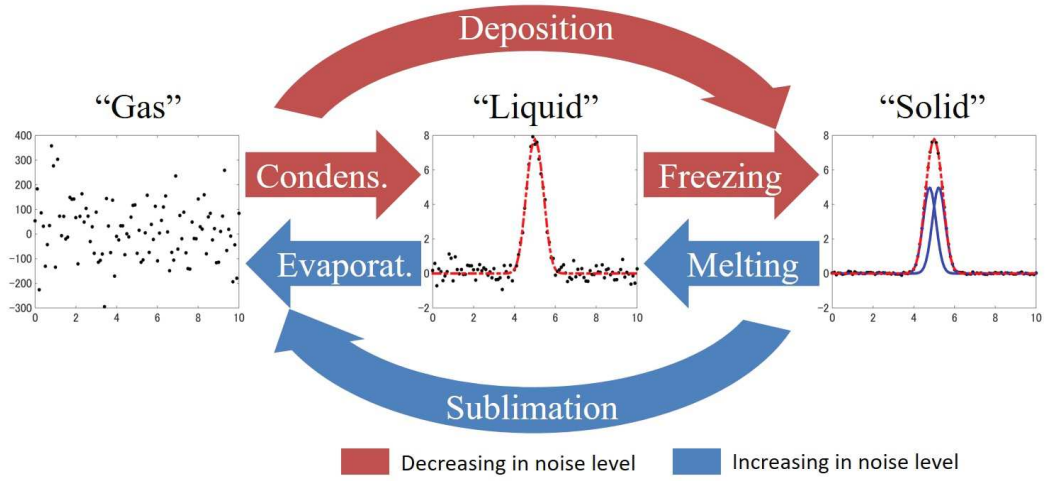

Figure 1.3: Phase transitions in statistical estimation.

noise cannot be distinguished. There is a limit of estimation by way of Bayesian spectral deconvolution, which two individual signals cannot be distinguished, i.e. the "degeneracy" or not. We elucidate these limits are uniformly explained as phase transitions in statistical estimation by reformulating Bayesian spectral deconvolution.

1.3 Structure of this thesis

This thesis consists of five chapters. In Chapter 2, we formulate the inverse problem that calculate energy levels from an observed spectrum, a modification of Bayesian spectral deconvolution, and show a demonstration. In Chapter 3, we develop a mathematical foundation of phase transitions in statistical estimation based on Watanabe's formulation and show that the "degeneracy" is a phase transition in statistical estimation. In Chapter 4, we clarify the measurement limit of dispersive spectrometers with respect to the measurement time interval. In Chapter 5, we conclude this thesis.

Chapter 2

Modification of Bayesian spectral deconvolution

2.1 Introduction

It is generally a difficult problem to distinguish each peak from noisy spectra with overlapping peaks. The simplest solution is least-squares fitting by a gradient method [56]. This type of method has a drawback in that fitting parameters are often trapped at a local minimum or a saddle whenever there is another global minimum in the parameter space. Moreover, the number of peaks is not always known in practice. Bayesian inference, by using a Markov chain Monte Carlo (MCMC) method, provides a superior solution [46–50,57–60]. Although the Bayesian framework enables us to estimate the number of peaks, MCMC methods generally have the limitation of local minima and saddles. Nagata et al. reported [46] that the exchange Monte Carlo method [61] (or parallel tempering [62]) can prevent local minima or saddles efficiently and provide a more accurate estimation than the reversible jump MCMC method [63] and its extension [64].

We constructed a Bayesian framework for estimating both the noise variance and the number of peaks from spectra with white Gaussian noise by expanding the previous framework by Nagata et al. [46]. The noise variance and the number of peaks are respectively estimated by hyperparameter optimization and model selection. These estimations are carried out by maximizing a function called the marginal likelihood [43–45], which is a conditional probability of observed data given the noise variance and the number of peaks in our framework. We provide a straightforward and efficient scheme that calculates this bivariate function by using the exchange Monte Carlo method and the multiple histogram method [65,

66]. We also demonstrated our framework through simulation. We show that estimating both the noise variance and the number of peaks prevents overfitting, overpenalizing, and misunderstanding the precision of parameter estimation.

2.2 Framework

2.2.1 Models

An observed spectrum $y \in \mathbb{R}$ is represented by the sum $f(x; w)$ of single peaks $\phi_k(x; \mu_k, \rho_k)$ and additive noise ε as

$$y = f(x; w) + \varepsilon, \tag{2.1}$$

$$f(x; w) := \sum_{k=1}^{K} a_k \phi_k(x; \mu_k, \rho_k), \tag{2.2}$$

$$\phi_k(x; \mu_k, \rho_k) := \exp\left[-\frac{\rho_k}{2}(x - \mu_k)^2\right], \tag{2.3}$$

where $x \in \mathbb{R}$ denotes energy, frequency, or wave number depending on the case. The parameter set is $w := \{a_k, \mu_k, \rho_k\}_{k=1}^{K}$, where $a_k \geq 0$, $\mu_k \in \mathbb{R}$, and $\rho_k^{-1/2} (\rho_k \geq 0)$ for each k are respectively the intensity, energy level, and peak width. The Gaussian function $\phi_k(x)$ for each k should be replaced with other parametric functions, such as the Lorentzian or Voigt function, depending on the case [42, 67]. If the peaks $\phi_k(x)$ are symmetric functions for all k (i.e., their values depend only on the distance from each center), the function $f(x; w)$ is called a radial basis function network in neural networks and related fields [46, 68]. This is the junction of the spectral data analysis and singular learning theory [69]. If the additive noise ε is assumed to be a zero-mean Gaussian with variance $b^{-1} \geq 0$, the statistical model of the observed spectrum is represented by a conditional probability as

$$p(y \mid x, w, b) := \sqrt{\frac{b}{2\pi}} \exp\left\{-\frac{b}{2}[y - f(x; w)]^2\right\}, \tag{2.4}$$

where y is taken as a random variable. This Gaussian distribution $p(y \mid x, w, b)$ is valid if the thermal noise is dominant. The parameter set w is also regarded as a random variable from the Bayesian viewpoint. The probability density function of w, called the *prior* density, is heuristically modeled as

$$\varphi(w \mid K) := \prod_{k=1}^{K} \varphi(a_k)\, \varphi(\mu_k)\, \varphi(\rho_k), \tag{2.5}$$

$$\varphi(a_k) := \kappa \exp(-\kappa a_k), \tag{2.6}$$

$$\varphi(\mu_k) := \sqrt{\frac{\alpha}{2\pi}} \exp\left[-\frac{\alpha}{2}(\mu_k - \mu_0)^2\right] \tag{2.7}$$

$$\varphi(\rho_k) := \nu \exp(-\nu\rho_k), \tag{2.8}$$

where $\kappa > 0$, $\mu_0 \in \mathbb{R}$, $\alpha > 0$, and $\nu > 0$ are hyperparameters. This prior density modeling is a special case of that by Nagata et al. [46]. Equation (2.6) promotes the sparsity of a_k. Equation (2.7) is regarded as an almost flat prior density if α is sufficiently small. These prior density models can be replaced with any other model without loss of generality in our framework.

2.2.2 Bayesian formulation

The conditional probability density function of w given samples $D := \{X_i, Y_i\}_{i=1}^n$, set as $X_1 < X_2 < \cdots < X_n$ for the sake of convenience, is represented by Bayes' theorem as

$$p(w \mid D, K, b) = \frac{1}{Z_n(K,b)} \prod_{i=1}^n p(Y_i \mid X_i, w, b)\varphi(w \mid K) \tag{2.9}$$

$$= \frac{1}{\tilde{Z}_n(K,b)} \exp\left[-nbE_n(w)\right] \varphi(w \mid K), \tag{2.10}$$

$$Z_n(K,b) := \int dw \prod_{i=1}^n p(Y_i \mid X_i, w, b)\varphi(w \mid K) \tag{2.11}$$

$$= \left(\frac{b}{2\pi}\right)^{\frac{n}{2}} \tilde{Z}_n(K,b), \tag{2.12}$$

$$\tilde{Z}_n(K,b) := \int dw \exp\left[-nbE_n(w)\right] \varphi(w \mid K), \tag{2.13}$$

$$E_n(w) := \frac{1}{2n} \sum_{i=1}^n \left[Y_i - f(X_i; w)\right]^2, \tag{2.14}$$

where the functions $p(w \mid D, K, b)$ and $Z_n(K,b)$ are respectively called the *posterior* density and marginal likelihood. Note that the function $Z_n(K,b) = p(\{Y_i\}_{i=1}^n \mid \{X_i\}_{i=1}^n K, b)$ is a probability density but $\tilde{Z}_n(K,b)$ is not. Bayes free energy $F_n(K,b)$ is defined as

$$F_n(K,b) := -\log Z_n(K,b) \tag{2.15}$$

$$= b\tilde{F}_n(K,b) - \frac{n}{2}(\log b - \log 2\pi), \tag{2.16}$$

$$\tilde{F}_n(K,b) := -\frac{1}{b} \log \tilde{Z}_n(K,b). \tag{2.17}$$

Note that Nagata et al. regarded $b\tilde{F}_n(K,b)$ as Bayes free energy for the sake of convenience [46] since the noise variance is treated as a known constant. We also assume the case in which there are no peaks as $K = 0$ (see Appendix A). In terms of the empirical Bayes (or type II maximum likelihood) approach [43–45], empirical Bayes estimators of K and b are given by

$$(\hat{K}, \hat{b}) := \arg\max_{K,b} Z_n(K,b) \tag{2.18}$$

$$= \arg\min_{K,b} F_n(K,b). \tag{2.19}$$

The hierarchical Bayes approach [70] is also tractable in our framework (see Appendix B). The partial derivative of $F_n(K,b)$ with respect to the variable b is obtained as

$$\frac{\partial F_n}{\partial b} = n\left[\langle E_n(w)\rangle_b - \frac{1}{2b}\right], \tag{2.20}$$

where $\langle Q\rangle_b$ denotes the posterior mean of an arbitrary quantity $Q \in \mathbb{R}$ over $p(w \mid D, K, b)$. If $b = \hat{b}$ is a stationary point of $F_n(K,b)$, then the following equation is satisfied:

$$\langle E_n(w)\rangle_{\hat{b}} = \frac{1}{2\hat{b}}. \tag{2.21}$$

The Bayes estimator of w is given by $\hat{w} := \{\langle a_k\rangle_{\hat{b}}, \langle \mu_k\rangle_{\hat{b}}, \langle \rho_k\rangle_{\hat{b}}\}_{k=1}^{\hat{K}}$ with the standard deviation $\sqrt{\langle Q'^2\rangle_{\hat{b}} - \langle Q'\rangle_{\hat{b}}^2}$ for each parameter $Q' \in w$ if $\hat{K} > 0$. However, $(\hat{K}, \hat{b})$ cannot be derived in this case since $F_n(K,b)$ and $\langle E_n(w)\rangle_b$ are analytically intractable for our model.

2.3 Algorithm

2.3.1 Exchange Monte Carlo method

In practice, we calculate $F_n(K,b)$ and $\langle E_n(w)\rangle_b$ by using the exchange Monte Carlo method, which efficiently enables sampling from $p(w \mid D, K, b)$ at $b \in \{b_l\}_{l=1}^L$ without knowing $Z_n(K,b)$ or $F_n(K,b)$. The target density is a joint probability density as

$$p\left(\{w_l\}_{l=1}^L \mid D, K, \{b_l\}_{l=1}^L\right) := \prod_{l=1}^L p(w_l \mid D, K, b_l), \tag{2.22}$$

where w_l is the parameter set at b_l. Each density $p(w_l \mid D, K, b_l)$ is called a *replica*. Sequence $\{b_l\}_{l=1}^L$ is set as $0 = b_1 < b_2 < \cdots < b_L$ for the sake of convenience. Note

that the variable b is replaced with the inverse temperature β of Nagata et al.'s formulation [46]. The variable b works as quasi-inverse temperature and varies the substantial support of the posterior density $p(w \mid D, K, b)$. The state exchange between high- and low-temperature replicas enables the escape from local minima or saddles in the parameter space. The sampling procedure includes the two following steps.

- State update in each replica
  Simultaneously and independently update state w_l subject to $p(w_l \mid D, K, b_l)$ using the Metropolis algorithm [71].

- State exchange between neighboring replicas
  Exchange states w_l and w_{l+1} at every step subject to the probability $u(w_{l+1}, w_l, b_{l+1}, b_l)$ as

$$u(w_{l+1}, w_l, b_{l+1}, b_l) := \min\left[1, v(w_{l+1}, w_l, b_{l+1}, b_l)\right], \tag{2.23}$$

$$v(w_{l+1}, w_l, b_{l+1}, b_l) := \frac{p(w_{l+1} \mid D, K, b_l)p(w_l \mid D, K, b_{l+1})}{p(w_l \mid D, K, b_l)p(w_{l+1} \mid D, K, b_{l+1})} \tag{2.24}$$

$$= \exp\left\{n(b_{l+1} - b_l)[E_n(w_{l+1}) - E_n(w_l)]\right\}, \tag{2.25}$$

  where Eq. (2.23) ensures a detailed balance condition.

A straightforward way of computing $\tilde{F}_n(K, b_l)$ via the exchange Monte Carlo method is bridge sampling [72, 73], in which $\tilde{F}_n(K, b_l)$ is expressed as

$$\tilde{F}_n(K, b_l) = -\frac{1}{b_l} \log \prod_{l'=1}^{l-1} \frac{\tilde{Z}(K, b_{l'+1})}{\tilde{Z}(K, b_{l'})} \tag{2.26}$$

$$= -\frac{1}{b_l} \sum_{l'=1}^{l-1} \log\langle \exp[-n(b_{l'+1} - b_{l'})E_n(w_{l'})]\rangle_{b_{l'}}, \tag{2.27}$$

where $\langle Q_l \rangle_{b_l}$ for the arbitrary quantity $Q_l \in \mathbb{R}$ at the l^{th} replica is approximated by the mean of an MCMC sample $\{Q_{l,m}\}_{m=1}^{M_l}$ as

$$\langle Q_l \rangle_{b_l} = \frac{1}{M_l} \sum_{m=1}^{M_l} Q_{l,m}. \tag{2.28}$$

However, $\hat{b}$ is not easy to accurately calculate using only the above scheme since $\{b_l\}_{l=1}^{L}$ is a discrete set, whereas b is a continuous variable.

2.3.2 Multiple histogram method

We interpolate $\{F_n(K, b_l)\}_{l=1}^L$ or $\{\langle E_n(w)\rangle_{b_l}\}_{l=1}^L$ with respect to $b = b' \in (b_l, b_{l+1})$ for any l via the multiple histogram method. The density of states is defined and estimated by

$$g(E; K) := \int dw \delta[E - E_n(w)]\varphi(w \mid K) \tag{2.29}$$

$$= \frac{\sum_{l=1}^L N_l(E)}{\sum_{l'=1}^L M_{l'} \tilde{Z}_n(K, b_{l'})^{-1} \exp(-n b_{l'} E)}, \tag{2.30}$$

then we obtain

$$\tilde{Z}_n(K, b) = \int dE g(E; K) \exp(-nbE) \tag{2.31}$$

$$= \sum_{l=1}^L \sum_{m=1}^{M_l} \frac{1}{\sum_{l'=1}^L M_{l'} \tilde{Z}_n(K, b_{l'})^{-1} \exp\left[n(b - b_{l'})E_{l,m}\right]}, \tag{2.32}$$

where $N_l(E)dE$ and $E_{l,m}$ are respectively the histogram of $E \geq 0$ at the l^{th} replica and the value of E at the m^{th} snapshot of the l^{th} replica in an MCMC simulation, i.e., $\int dE N_l(E) = M_l$. The values of $\{\tilde{Z}_n(K, b_l)\}_{l=1}^L$ are determined self-consistently by iterating Eq. (2.32) with $b = b_l$. We take $\exp[-b_l \tilde{F}_n(K, b_l)]$ computed via Eq. (2.27) as the initial values for the sake of convenience. Given $\{\tilde{Z}_n(K, b_l)\}_{l=1}^L$, we then calculate $\tilde{Z}_n(K, b)$ as $b = b'$ via Eq. (2.32) again. The above procedure can be appropriately generalized to treat multidimensional histograms such as $N_l(E, Q)dEdQ$ [74]. Then, the posterior mean of an arbitrary quantity is calculated as

$$\langle Q\rangle_b = \frac{1}{\tilde{Z}_n(K, b)} \sum_{l=1}^L \sum_{m=1}^{M_l} \frac{Q_{l,m}}{\sum_{l'=1}^L M_{l'} \tilde{Z}_n(K, b_{l'})^{-1} \exp\left[n(b - b_{l'})E_{l,m}\right]}, \tag{2.33}$$

where $Q_{l,m}$ is the value of Q at the m^{th} snapshot of the l^{th} replica in an MCMC simulation. We calculate $\langle E_n(w)\rangle_b$ via Eq. (2.33) and solve Eq. (2.21) numerically by the bisection method. Then, $\hat{w}$ with the standard deviation of each parameter is also calculated via Eq. (2.33). The posterior density of arbitrary quantities can also be interpolated with respect to $b = b'$ in the same way (see Appendix C).

2.4 Simulation

We demonstrated how efficient our framework is through simulation in which the same synthetic data as used by Nagata et al. [46] were used. The synthetic data

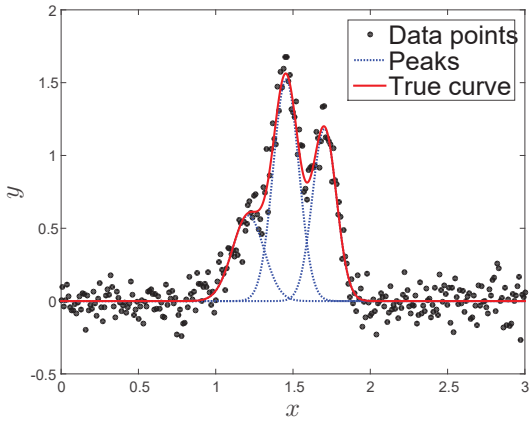

Figure 2.1: Synthetic data. The horizontal and vertical axes respectively represent the input x and output y. The black dots show synthetic data $D = \{X_i, Y_i\}_{i=1}^n$. The red solid line and blue dotted ones respectively show the true curve $y = f(x; w_0)$ and the Gaussian peaks $y = \phi_k(x; {\mu_k}^*, {\rho_k}^*)$.

$D = \{X_i, Y_i\}_{i=1}^n$ shown in Fig. 2.1 were generated from the true probability density as

$$q(y \mid x, w_0, b_0) := \sqrt{\frac{b_0}{2\pi}} \exp\left\{-\frac{b_0}{2}[y - f(x; w_0)]^2\right\}, \tag{2.34}$$

where $b_0 > 0$ and $w_0 := \{{a_k}^*, {\mu_k}^*, {\rho_k}^*\}_{k=1}^{K_0}$ are respectively the true inverse noise variance and true parameter set, as in Tables 2.1 and 2.2. The inputs $\{X_i\}_{i=1}^n$ were linearly spaced in the interval $[X_1, X_n] = [0, 3]$ with spectral resolution $\Delta x = 0.01$, where the number of samples was $n = 301$. The sequence $\{b_l\}_{l=2}^L$ were logarithmically spaced in the interval $[nb_2, nb_L] = [10^{-4}, 10^8]$, where the number of replicas was $L = 400$. The model size K was set as integers from 0 to 5. The hyperparameters were $\kappa = 1.7$, $\mu_0 = 1.5$, $\alpha = 0.4$, and $\nu = 0.01$ in the heuristics. The total number of MCMC sweeps was 100,000 including 50,000 burn-in sweeps: an MCMC sample $\{w_{l,m}\}_{m=1}^{M_l}$ of size $M_l = 50,000$ for every b_l was obtained. The estimators are listed in Tables 2.1 and 2.2, where ρ_k was converted into an inverse square-root scale for comparison.

First, we discuss how to estimate both the noise variance and the number of peaks. (A) Bayes free energy and (B) the posterior mean of the mean square error are shown in Fig. 2.2. The horizontal axes represent b on a log scale. The colored solid lines show $F_n(K, b_l)$ calculated via Eq. (2.27) for each K in (A) and $\langle E_n(w)\rangle_{b_l}$ calculated via Eq. (2.28) for each K on a log scale in (B). The three lines of $K \geq 3$ almost overlap in (A-1) and (B-1), whose enlarged views around

Table 2.1: Number of peaks and inverse noise variance.

	K	b
Estimated	3	1.02941×10^2
True	3	1.0000×10^2

Table 2.2: Parameters of each Gaussian peak.

		a_k	μ_k	${\rho_k}^{-1/2}$
Mode 1	Estimated	0.579 ± 0.054	1.257 ± 0.040	0.14413 ± 0.02571
($k = 1$)	True	0.587	1.210	0.10223
Mode 2	Estimated	1.351 ± 0.152	1.461 ± 0.004	0.0760612 ± 0.0060438
($k = 2$)	True	1.522	1.455	0.0825244
Mode 3	Estimated	1.160 ± 0.048	1.703 ± 0.004	0.0817504 ± 0.0040759
($k = 3$)	True	1.183	1.703	0.0779755

the black circles are respectively shown in (A-2) and (B-2). The colored markers in (A-2) and (B-2) respectively indicate $F_n(K, b_l)$ as in (A-1) and $\langle E_n(w) \rangle_{b_l}$ as in (B-1). The colored dotted lines in (A-2) and (B-2) respectively indicate the interpolated values calculated via Eqs. (2.32) and (2.33). The gray solid lines in (B) show the function $1/2b$. The vertical black dashed lines and vertical black dash-dotted ones respectively show the true value $b = b_0$ and the estimated value $b = \hat{b}$. There is a minimum point of $F_n(K, b)$ depending on each value of K, i.e., the probability density $p(K, b \mid D)$ has a maximum at this point (see Appendix B). In this case, Eq. (2.21) holds at the intersection of the purple dotted line and the gray solid line shown in (B-2).

Second, we discuss the validity of our framework. The dependence on b in the model selection is shown in Fig. 2.3. The horizontal axis represents b on a log scale. The colored markers show the estimated model size $\hat{K}_b$ that minimizes $F_n(K, b_l)$ for each b_l as

$$\hat{K}_b := \arg\min_K F_n(K, b_l) \tag{2.35}$$

$$= \arg\min_K \tilde{F}_n(K, b_l). \tag{2.36}$$

Note that $\hat{K}_{b_0} = \arg\min_K \tilde{F}_n(K, b_0)$ is regarded as the optimal number of peaks in Nagata et al.'s framework [46]. The vertical black dashed line and the vertical black dash-dotted one respectively show the true value $b = b_0$ and the estimated

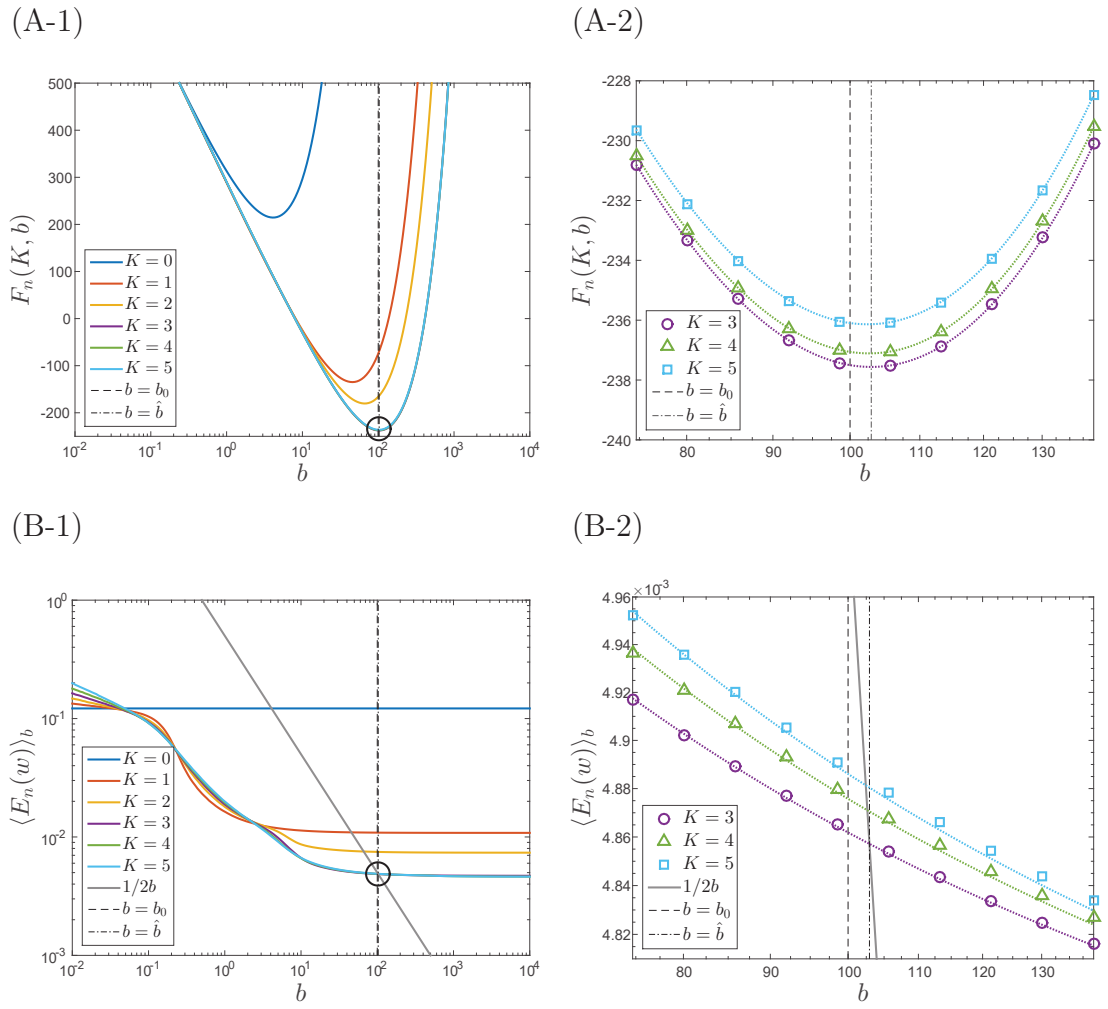

Figure 2.2: (A) Bayes free energy and (B) posterior mean of mean square error. The horizontal axes represent b on a log scale. The colored solid lines show $F_n(K, b_l)$ for each K in (A) and $\langle E_n(w)\rangle_{b_l}$ for each K on a log scale in (B). The three lines of $K \geq 3$ almost overlap in (A-1) and (B-1) whose enlarged views around black circles are respectively shown in (A-2) and (B-2). The colored markers in (A-2) and (B-2) respectively indicate $F_n(K, b_l)$ as in (A-1) and $\langle E_n(w)\rangle_{b_l}$ as in (B-1). The colored dotted lines in (A-2) and (B-2) indicate the interpolated values. The gray solid lines in (B) show the function $1/2b$. The vertical black dashed lines and vertical black dash-dotted ones respectively show the true value $b = b_0$ and the estimated value $b = \hat{b}$.

value $b = \hat{b}$. Although $\hat{K}_b$ for each value of b depends on the noise realization, as Nagata et al. showed in the case of $b = b_0$ [46], $\hat{K}_b$ also changes depending on the value of b. There is a rough trend, explained by the asymptotic form of $\tilde{F}_n(K, b)$, in which $\hat{K}_b$ becomes larger as b increases. If the sample size n is sufficiently large,

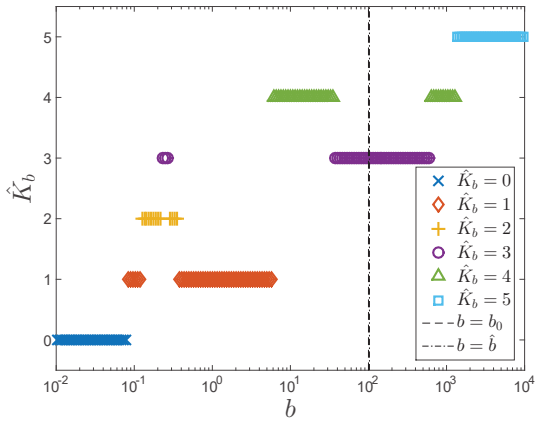

Figure 2.3: Dependence of model selection on b. The horizontal axis represents b on a log scale. The estimated model size $\hat{K}_b$ that minimizes $F_n(K, b)$ for each b is plotted as colored marker. The vertical black dashed line and the vertical black dash-dotted one respectively show the true value $b = b_0$ and the estimated value $b = \hat{b}$.

$\tilde{F}_n(K, b)$ is expressed as

$$\tilde{F}_n(K, b) = nE_n(w_0) + \frac{\lambda}{b} \log n + \frac{1}{b} O_p(\log \log n), \tag{2.37}$$

where w_0 is the parameter set that minimizes the Kullback–Leibler divergence of a statistical model from a true distribution, and $\lambda > 0$ is a rational number called the real log canonical threshold (RLCT) [52, 53]. The RLCT is determined by the pair of a statistical model and true distribution, and the ones determined by Eqs. (2.4) and (2.34) are clarified for several cases of (K, K_0) with $b = b_0$ [69]. The values $E_n(w_0)$ and λ respectively become larger and smaller as K increases. The term $nE_n(w_0)$ dominantly works for model selection for large b: overfitting occurs. The term $\lambda \log n$ dominantly works for small b: overpenalizing occurs. A moderate model is estimated under the moderate value of b. Estimating the optimal value of b is indispensable, and this result shows the validity of our framework.

Finally, we discuss the validity of our framework from another viewpoint. (A) The posterior mean of μ_k, (B) the posterior standard deviation of μ_k, and (a-d) the marginal posterior distribution of μ_k when $K = K_0 = 3$ are shown in Fig. 2.4. The horizontal axes in (A-B) represent b on a log scale. The colored solid lines show $\langle \mu_k \rangle_{b_l}$ for each k in (A) and $2\sqrt{\langle {\mu_k}^2 \rangle_{b_l} - \langle \mu_k \rangle_{b_l}^2}$ for each k in log scale in (B). These values were calculated via Eq. (2.28). The identification of mode k was reassigned by sorting the MCMC sample $\{\mu_{k,l,m}\}_{k=1}^3$ into $\mu_{1,l,m} < \mu_{2,l,m} < \mu_{3,l,m}$ for each l and m in light of the exchange symmetry. The vertical black dashed

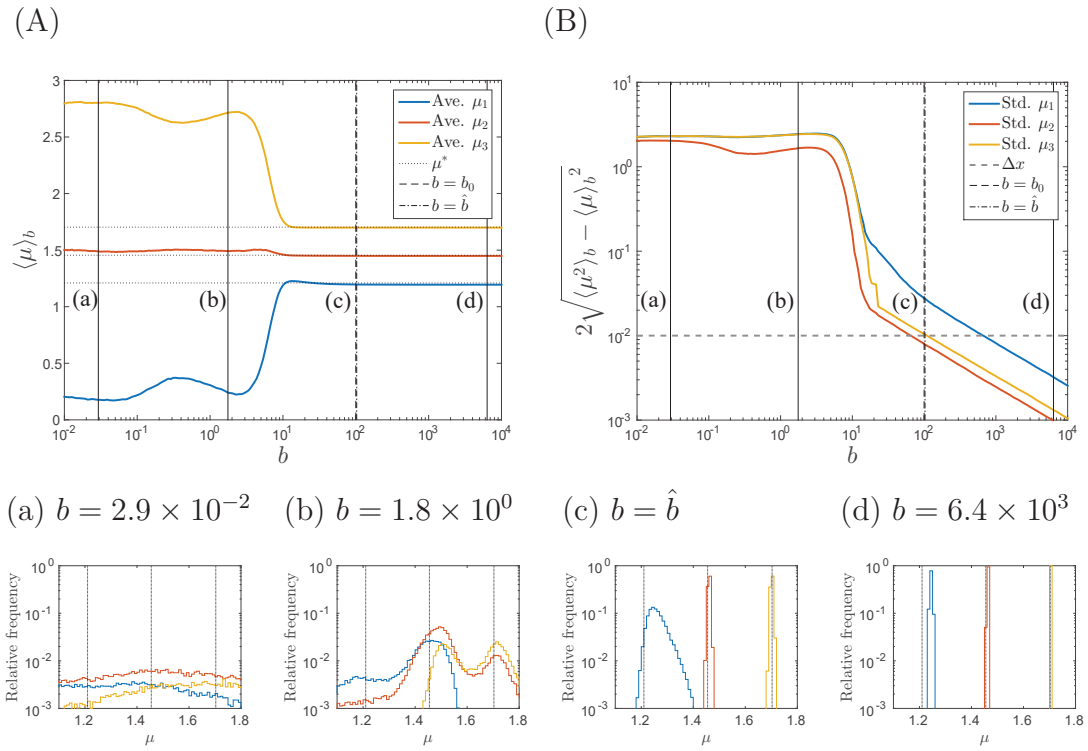

Figure 2.4: (A) Posterior mean of μ_k, (B) posterior standard deviation of μ_k, and (a-d) marginal posterior distribution of μ_k when $K = K_0 = 3$. The horizontal axes in (A-B) represent b on a log scale. The colored solid lines show $\langle \mu_k \rangle_{b_l}$ for each k in (A) and $2\sqrt{\langle {\mu_k}^2 \rangle_{b_l} - \langle \mu_k \rangle_{b_l}^2}$ for each k on a log scale in (B). The vertical black dashed lines and the vertical black dash-dotted ones respectively show the true value $b = b_0$ and the estimated value $b = \hat{b}$. The horizontal black dotted lines in (A) show the true value ${\mu_k}^*$ for each k and the horizontal gray dashed line in (B) shows Δx. The vertical black solid lines in (A-B) correspond to each value of b in (a-d). The histograms (a-d) of μ_k show the marginal posterior distribution of μ_k for each b, where the coloring for each μ_k follows that in (A-B). The horizontal axes in (a-d) represent μ_k, and the vertical ones represent relative frequency on a log scale. The vertical black dotted lines also show the true value ${\mu_k}^*$ for each k, as in (A).

lines and the vertical black dash-dotted ones respectively show the true value $b = b_0$ and the estimated value $b = \hat{b}$. The horizontal black dotted lines in (A) show the true value μ_k^* for each k and the horizontal gray dashed line in (B) shows the spectral resolution Δx. The vertical black solid lines in (A-B) correspond to each value of b in (a-d). The relative frequency histograms (a-d) show the marginal

posterior probability of μ_k for each bin $[X_i, X_{i+1}]$ and b as follows:

$$P(X_i \leq \mu_k \leq X_{i+1} \mid D, K, b) = \int_{X_i}^{X_{i+1}} d\mu_k p(\mu_k \mid D, K, b), \tag{2.38}$$

$$p(\mu_k \mid D, K, b) = \int dw' p(w \mid D, K, b) \tag{2.39}$$

$$= \frac{\tilde{z}_n(K, b, \mu_k)\varphi(\mu_k)}{\tilde{Z}_n(K, b)}, \tag{2.40}$$

$$\tilde{z}_n(K, b, \mu_k) := \int dw' \exp\left[-nbE_n(w'; \mu_k)\right] \varphi(w' \mid K), \tag{2.41}$$

where $w' := w \backslash \{\mu_k\}$ and $\varphi(w' \mid K) := \varphi(w \mid K)/\varphi(\mu_k)$. $E_n(w'; \mu_k)$ indicates the function $E_n(w)$ given the value μ_k. The histograms (a), (b), and (d) were respectively constructed using the MCMC sample $\{\mu_{k,l,m}\}_{m=1}^{M_l}$ as $b = 2.925210 \times 10^{-2}, 1.758132 \times 10^{0}, 6.350977 \times 10^{3}$ for each k. Histogram (c) was calculated via Eq. (A.17) for each k (see Appendix C). The coloring of the histogram for each k follows that in (A-B). The horizontal axes in (a-d) represent μ_k, and the vertical ones represent relative frequency on a log scale. The vertical black dotted lines in (a-d) show the true value μ_k^* for each k, as in (A). $\langle \mu_k \rangle_{b_l}$ and $2\sqrt{\langle {\mu_k}^2 \rangle_{b_l} - \langle \mu_k \rangle_{b_l}^2}$ respectively change depending on b, where the changes in the support of the posterior density correspond. These changes are considerable around $b = 10^1$, where $\langle \mu_k \rangle_b$ for each k asymptotically approaches the true value ${\mu_k}^*$ from this region and $2\sqrt{\langle {\mu_k}^2 \rangle_b - \langle \mu_k \rangle_b^2}$ for each k monotonically decreases from the same region. The marginal posterior densities of μ_1, μ_2, and μ_3 overlap and are unidentifiable if b is smaller than around 10^1. Otherwise, they are separated and identifiable. $2\sqrt{\langle {\mu_2}^2 \rangle_b - \langle \mu_2 \rangle_b^2}$ is smaller than Δx as (c) $b = \hat{b}$: a kind of super-resolution. This effect is based on the same principle as super-resolution microscopy techniques [75, 76]. $2\sqrt{\langle {\mu_k}^2 \rangle_b - \langle \mu_k \rangle_b^2}$ for each k is also smaller than Δx as (d) $b > \hat{b}$, whereas the support of μ_1 does not cover the true value μ_1^*: outside the confidence interval. An appropriate setting of b provides an appropriate precision of parameter estimation. Estimating the optimal value of b is indispensable even if the true model size K_0 is known; thus, this result also shows the validity of our framework.

2.5 Discussion

We constructed a framework that enables the dual estimation of the noise variance and the number of peaks and demonstrated the effectiveness of our framework

through simulation. We also warned that there are the risks of overfitting, overpenalizing, and misunderstanding the precision of parameter estimation without the estimation of the noise variance. Our framework is an extension of Nagata et al.'s framework and is versatile and applicable to not only spectral deconvolution but also any other nonlinear regression with hierarchical statistical models.

Our framework is also considered as a learning scheme in radial basis function networks. However, the goal of spectral deconvolution is not to predict any future data, which is the goal of most other learning tasks, but to identify the true model since spectral deconvolution is an inverse problem of physics. This is the reason why we do not adopt the Bayes generalization error but adopt the Bayes free energy for hyperparameter optimization and model selection. The Akaike information criterion (AIC) [77] and Bayesian information criterion (BIC) [54], which are respectively approximations of the generalization error and Bayes free energy, do not hold for hierarchical models such as radial basis function networks: the widely applicable information criterion (WAIC) [78] and widely applicable Bayesian information criterion (WBIC) [79] generally hold for any statistical model. If the noise variance is unknown, these criteria do not lead to computational reduction since the value of the noise variance needs to be estimated, as discussed in Sect. 2.4. The example we gave is classified as an unrealizable and singular (or regular) case [80], which is a difficult problem. On the other hand, the example Nagata et al. gave [46] is classified as a realizable and singular (or regular) case, which is a relatively easy problem. Statistical hypothesis testing does not hold for a singular case. Our scheme is also valid and sophisticated from the viewpoint of statistics.

Chapter 3

Phase transitions in statistical estimation

第 3 章は雑誌掲載が予定される内容を含むため，インターネット公表できません.

Chapter 4

Measurement limit of dispersive spectrometers

第４章は雑誌掲載が予定される内容を含むため，インターネット公表できません.

Chapter 5

Conclusions

We have focused on the problem of identifying by the observed spectrum whether the energy level is degenerate or not and studied effects of measurement noise on Bayesian spectral deconvolution. The contributions of this thesis are as follows:

1. We have modified the conventional Bayesian spectral deconvolution to enable estimating the noise variance as Bayesian inference originally does so, and shown the significance of estimating the noise variance and the number of peaks simultaneously.

2. We have developed a mathematical foundation of phase transitions in statistical estimation by introducing Bayes specific heat and its scaling law.

3. We have found a phenomenon that, in Bayesian spectral deconvolution, the estimated values of the energy levels are "degenerate", affected by the measurement noise, and clarified that this phenomenon is a first-order phase transition in statistical estimation.

4. We have derived the extrapolation formula of Bayes specific heat on the dispersive spectroscopic measurement, and clarified the phase transitions, which indicate the measurement limit, with respect to the measurement time interval.

5. We have made a proposition that the degeneracy in the measured system and the "degeneracy" of the estimated value affected by measurement noise are essentially nonidentifiable.

These contributions possibly give an impact not just on physics but also on statistics and the related research field, such as artificial intelligence, neural networks, and machine learning.

Our findings also could contribute to the design of experiment by way of two types of strategy. One way is to replace only the forward problem with ab initio or model calculations. By integrating these calculations into the simulation of the indirect measurement, we can predict the measurement limit before the experiment. The other way is to carry out a pre-experiment as noiseless as possible. By using a noiseless spectrum for an indirect measurement, we can get the information of the measurement limit as the transition point. Of course, there is no way to know whether what we observe is "degenerate" or not in cases of actual measurements. But what is observed by the best measurement is surely our scientific truth. At least the corresponding value of Bayes specific heat indicate the state of measurement. It is no exaggeration to say that this means a "measurement" of a measurement.

References

[1] J. M. Silverman, M. Ganeshalingam, W. Li, and A. V. Filippenko: Monthly Notices of the Royal Astronomical Society **425** (2012) 1889.

[2] S. Blondin, K. S. Mandel, and R. P. Kirshner: Astronomy & Astrophysics **526** (2011) A81.

[3] W. Collmar, M. Böttcher, T. Krichbaum, I. Agudo, E. Bottacini, M. Bremer, V. Burwitz, A. Cuccchiara, D. Grupe, and M. Gurwell: Astronomy & Astrophysics **522** (2010) A66.

[4] G. Aldering, P. Antilogus, C. Aragon, C. Baltay, S. Bongard, C. Buton, M. Childress, N. Chotard, Y. Copin, E. Gangler, et al.: Astronomy & Astrophysics **500** (2009) L17.

[5] V. Larionov, S. Jorstad, A. Marscher, C. Raiteri, M. Villata, I. Agudo, M. Aller, A. Arkharov, I. Asfandiyarov, U. Bach, et al.: Astronomy & Astrophysics **492** (2008) 389.

[6] G. Vedrenne, J.-P. Roques, V. Schönfelder, P. Mandrou, G. Lichti, A. Von Kienlin, B. Cordier, S. Schanne, J. Knödlseder, G. Skinner, et al.: Astronomy & Astrophysics **411** (2003) L63.

[7] A. Satapathy, S. T. Gadge, E. N. Kusumawati, K. Harada, T. Sasaki, D. Nishio-Hamane, and B. M. Bhanage: Catalysis Letters **145** (2015) 824.

[8] S. T. Gadge, E. N. Kusumawati, K. Harada, T. Sasaki, D. Nishio-Hamane, and B. M. Bhanage: Journal of Molecular Catalysis A: Chemical **400** (2015) 170.

[9] S. Ghosh, S. S. Acharyya, T. Sasaki, and R. Bal: Green Chemistry **17** (2015) 1867.

[10] M. A. Bhosale, D. Ummineni, T. Sasaki, D. Nishio-Hamane, and B. M. Bhanage: Journal of Molecular Catalysis A: Chemical **404** (2015) 8.

[11] M. Tada, R. Bal, T. Sasaki, Y. Uemura, Y. Inada, S. Tanaka, M. Nomura, and Y. Iwasawa: The Journal of Physical Chemistry C **111** (2007) 10095.

[12] M. Horio, T. Adachi, Y. Mori, A. Takahashi, T. Yoshida, H. Suzuki, L. Ambolode II, K. Okazaki, K. Ono, H. Kumigashira, et al.: Nature communications **7** (2016).

[13] T. Shimojima, K. Okazaki, and S. Shin: Journal of the Physical Society of Japan **84** (2015) 072001.

[14] K. Okazaki, Y. Ito, Y. Ota, Y. Kotani, T. Shimojima, T. Kiss, S. Watanabe, C.-T. Chen, S. Niitaka, T. Hanaguri, et al.: Scientific Reports **4** (2014).

[15] K. Konishi, T. Higuchi, J. Li, J. Larsson, S. Ishii, and M. Kuwata-Gonokami: Physical review letters **112** (2014) 135502.

[16] M. Sato, T. Higuchi, N. Kanda, K. Konishi, K. Yoshioka, T. Suzuki, K. Misawa, and M. Kuwata-Gonokami: Nature Photonics **7** (2013) 724.

[17] K. Yoshioka, Y. Morita, K. Fukuoka, and M. Kuwata-Gonokami: Physical Review B **88** (2013) 041201.

[18] J. Omachi, T. Suzuki, K. Kato, N. Naka, K. Yoshioka, and M. Kuwata-Gonokami: Physical review letters **111** (2013) 026402.

[19] K. Okazaki, Y. Ota, Y. Kotani, W. Malaeb, Y. Ishida, T. Shimojima, T. Kiss, S. Watanabe, C.-T. Chen, K. Kihou, et al.: Science **337** (2012) 1314.

[20] K. Okazaki, Y. Ito, Y. Ota, Y. Kotani, T. Shimojima, T. Kiss, S. Watanabe, C.-T. Chen, S. Niitaka, T. Hanaguri, et al.: Physical review letters **109** (2012) 237011.

[21] K. Konishi, M. Nomura, N. Kumagai, S. Iwamoto, Y. Arakawa, and M. Kuwata-Gonokami: Physical review letters **106** (2011) 057402.

[22] T. Ikeya, A. Sasaki, D. Sakakibara, Y. Shigemitsu, J. Hamatsu, T. Hanashima, M. Mishima, M. Yoshimasu, N. Hayashi, T. Mikawa, et al.: Nature protocols **5** (2010) 1051.

[23] A. Gautier, H. R. Mott, M. J. Bostock, J. P. Kirkpatrick, and D. Nietlispach: Nature structural & molecular biology **17** (2010) 768.

[24] D. Sakakibara, A. Sasaki, T. Ikeya, J. Hamatsu, T. Hanashima, M. Mishima, M. Yoshimasu, N. Hayashi, T. Mikawa, M. Wälchli, et al.: Nature **458** (2009) 102.

[25] E. B. Bertelsen, L. Chang, J. E. Gestwicki, and E. R. Zuiderweg: Proceedings of the National Academy of Sciences **106** (2009) 8471.

[26] H. J. Kim, S. C. Howell, W. D. Van Horn, Y. H. Jeon, and C. R. Sanders: Progress in nuclear magnetic resonance spectroscopy **55** (2009) 335.

[27] S. Hiller, R. G. Garces, T. J. Malia, V. Y. Orekhov, M. Colombini, and G. Wagner: Science **321** (2008) 1206.

[28] R. Sprangers, A. Velyvis, and L. E. Kay: Nature methods **4** (2007) 697.

[29] J. Fiaux, E. B. Bertelsen, A. L. Horwich, and K. Wüthrich: Nature **418** (2002) 207.

[30] T. Yabuki, T. Kigawa, N. Dohmae, K. Takio, T. Terada, Y. Ito, E. D. Laue, J. A. Cooper, M. Kainosho, and S. Yokoyama: Journal of biomolecular NMR **11** (1998) 295.

[31] S. Obata, K. Saiki, T. Taniguchi, T. Ihara, Y. Kitamura, and Y. Matsumoto: Journal of the Physical Society of Japan **84** (2015) 121012.

[32] T.-o. Terasawa and K. Saiki: Nature communications **6** (2015).

[33] D. Basov, M. Fogler, A. Lanzara, F. Wang, Y. Zhang, et al.: Reviews of Modern Physics **86** (2014) 959.

[34] B. H. Horgan, E. A. Cloutis, P. Mann, and J. F. Bell: Icarus **234** (2014) 132.

[35] L. C. Cheek and C. M. Pieters: American Mineralogist **99** (2014) 1871.

[36] R. G. Mayne, J. M. Sunshine, H. Y. McSween, T. J. McCoy, C. M. Corrigan, and A. Gale: Meteoritics & Planetary Science **45** (2010) 1074.

[37] J. M. Sunshine, C. M. Pieters, and S. F. Pratt: Journal of Geophysical Research: Solid Earth **95** (1990) 6955.

[38] H. Tanabe, A. Kuwahata, H. Oka, M. Annoura, H. Koike, K. Nishida, S. You, Y. Narushima, A. Balandin, M. Inomoto, et al.: Nuclear Fusion **53** (2013) 093027.

[39] K. Hill, M. Bitter, S. Scott, A. Ince-Cushman, M. Reinke, J. Rice, P. Beiersdorfer, M. Gu, S. Lee, C. Broennimann, et al.: The Review of scientific instruments **79** (2008) 10E320.

[40] A. Balandin and Y. Ono: The European Physical Journal D-Atomic, Molecular, Optical and Plasma Physics **17** (2001) 337.

[41] I. Condrea, E. Haddad, B. Gregory, and G. Abel: Physics of Plasmas (1994-present) **7** (2000) 3641.

[42] N. V. Tkachenko: *Optical spectroscopy: methods and instrumentations* (Elsevier, 2006).

[43] B. Efron and C. Morris: Journal of the American Statistical Association **68** (1973) 117.

[44] H. Akaike: Trabajos de estadística y de investigación operativa **31** (1980) 143.

[45] D. J. MacKay: Neural computation **4** (1992) 415.

[46] K. Nagata, S. Sugita, and M. Okada: Neural Networks **28** (2012) 82.

[47] P. Hong, H. Miyamoto, T. Niihara, S. Sugita, K. Nagata, J. M. Dohm, and M. Okada: Journal of Geology & Geophysics **5** (2016) 2.

[48] K. Hagino: Physical Review C **93** (2016) 061601.

[49] T. Kasai, K. Nagata, M. Okada, and T. Kigawa: Journal of Physics: Conference Series, Vol. 699, 2016, p. 012003.

[50] S. Murata, K. Nagata, M. Uemura, and M. Okada: to be published in Journal of the Physical Society of Japan .

[51] S. Tokuda, K. Nagata, and M. Okada: arXiv preprint arXiv:1607.07590 (2016).

[52] S. Watanabe: Neural Computation **13** (2001) 899.

[53] S. Watanabe: *Algebraic geometry and statistical learning theory* (Cambridge University Press, 2009), Vol. 25.

[54] G. Schwarz: The annals of statistics **6** (1978) 461.

[55] 渡辺澄夫: **ベイズ統計の理論と方法** (コロナ社, 2012).

[56] G. C. Allen and R. F. McMeeking: Analytica Chimica Acta **103** (1978) 73.

[57] R. Fischer and V. Dose: Bayesian Methods: With Applications to Science, Policy, and Official Statistics, 2001, pp. 145–154.

[58] S. G. Razul, W. Fitzgerald, and C. Andrieu: Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **497** (2003) 492.

[59] A. Masson, L. Poisson, M.-A. Gaveau, B. Soep, J.-M. Mestdagh, V. Mazet, and F. Spiegelman: The Journal of Chemical Physics **133** (2010) 054307.

[60] V. Mazet, S. Faisan, S. Awali, M.-A. Gaveau, and L. Poisson: Signal Processing **109** (2015) 193.

[61] K. Hukushima and K. Nemoto: Journal of the Physical Society of Japan **65** (1996) 1604.

[62] C. J. Geyer: Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface, 1991, pp. 156–163.

[63] P. J. Green: Biometrika **82** (1995) 711.

[64] A. Jasra, D. A. Stephens, and C. C. Holmes: Biometrika **94** (2007) 787.

[65] A. M. Ferrenberg and R. H. Swendsen: Physical Review Letters **63** (1989) 1195.

[66] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman: Journal of computational chemistry **13** (1992) 1011.

[67] R. Loudon: *The quantum theory of light* (Oxford University Press, 2000).

[68] D. S. Broomhead and D. Lowe: Complex Systems **2** (1988) 321.

[69] S. Tokuda, K. Nagata, and M. Okada: IPSJ Transactions on Mathematical Modeling and Its Applications **6** (2013) 117.

[70] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin: *Bayesian Data Analysis, Third Edition* (CRC Press, 2013).

[71] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller: The journal of chemical physics **21** (1953) 1087.

[72] X.-L. Meng and W. H. Wong: Statistica Sinica (1996) 831.

[73] A. Gelman and X.-L. Meng: Statistical science (1998) 163.

[74] M. Newman and G. Barkema: *Monte Carlo Methods in Statistical Physics* (Oxford University Press: New York, USA, 1999).

[75] E. Betzig: Optics letters **20** (1995) 237.

[76] E. Betzig, G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacino, M. W. Davidson, J. Lippincott-Schwartz, and H. F. Hess: Science **313** (2006) 1642.

[77] H. Akaike: IEEE transactions on automatic control **19** (1974) 716.

[78] S. Watanabe: Neural Networks **23** (2010) 20.

[79] S. Watanabe: The Journal of Machine Learning Research **14** (2013) 867.

[80] S. Watanabe: Journal of Physics: Conference Series, Vol. 233, 2010, p. 012014.

[81] D. J. MacKay: Maximum entropy and Bayesian methods, 1996, pp. 43–59.

Appendix A

Appendix in Chapter 2

A.1 Bayes free energy for no-peaks model

We define the function $f(x; w = \phi) = 0$ as $K = 0$, where ϕ is the empty set. The statistical model of the no-peaks spectrum and marginal likelihood are expressed as

$$p(y \mid x, w = \phi, b) = \sqrt{\frac{b}{2\pi}} \exp\left(-\frac{b}{2}y^2\right), \tag{A.1}$$

$$Z_n(K = 0, b) = \prod_{i=1}^{n} p(Y_i \mid X_i, w = \phi, b) \tag{A.2}$$

$$= \left(\frac{b}{2\pi}\right)^{\frac{n}{2}} \tilde{Z}_n(K = 0, b), \tag{A.3}$$

$$\tilde{Z}_n(K = 0, b) = \exp[-nbE_n(w = \phi)], \tag{A.4}$$

$$E_n(w = \phi) = \frac{1}{2n} \sum_{i=1}^{n} {Y_i}^2. \tag{A.5}$$

The main term of Bayes free energy and the posterior mean of the mean square error are also respectively expressed as

$$\tilde{F}_n(K = 0, b) = nE_n(w = \phi), \tag{A.6}$$

$$\langle E_n(w = \phi)\rangle_b = E_n(w = \phi), \tag{A.7}$$

where they can be calculated without any MCMC method.

A.2 Hierarchical Bayes approach

In Sect. 2.4, we adopted the empirical Bayes (or type II maximum likelihood) approach, in which K and b are estimated by the minimization of $F_n(K, b)$ (or

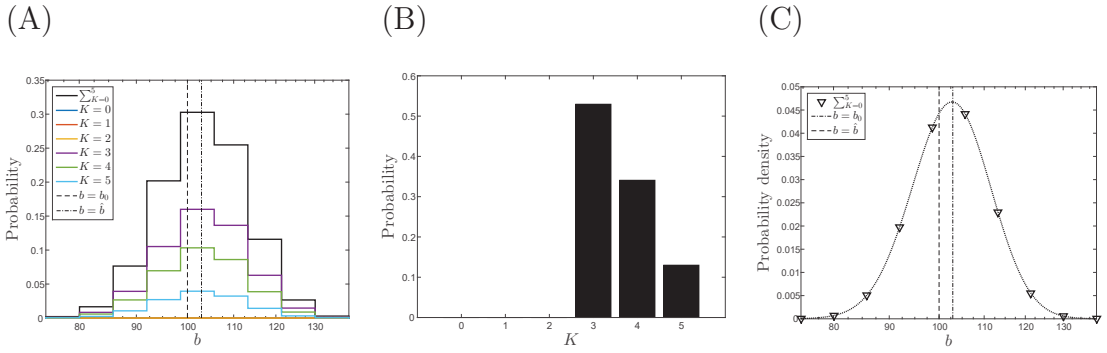

Figure A.1: (A) Joint probability of (K, b) and marginal probability of b, (B) marginal probability of K, and (C) marginal probability density of b. The horizontal axes represent b on a log scale. The colored stairstep graphs and the black one in (A) respectively show the joint probability $P(K, b_l \leq b \leq b_{l+1} \mid D)$ for each K and the marginal probability $P(b_l \leq b \leq b_{l+1} \mid D)$. The three colored graphs of $K < 3$ almost overlap in contrast to Fig. 2.2(A-1). The black bars in (B) show the marginal probability $P(K \mid D)$. The black markers and black dotted line in (C) respectively show the marginal probability density $p(b_l \mid D)$ and the interpolated values. The vertical black dashed lines and the vertical black dash-dotted ones respectively show the true value $b = b_0$ and the estimated value $b = \hat{b}$, as in Fig. 2.2.

the maximization of $Z_n(K, b)$). The hierarchical Bayes approach, which takes into account the posterior density of K and b, is also suitable for our framework. The prior density of K and b is set as $\varphi(K, b) = \varphi(K)\varphi(b)$, where $\varphi(K)$ is a discrete uniform distribution on the natural numbers $\{0, 1, 2, 3, 4, 5\}$ and $\varphi(b)$ is a continuous uniform distribution on the interval $[b_1, b_L]$. The joint posterior probability and marginal ones are expressed as

$$P(K, b_l \leq b \leq b_{l+1} \mid D) = \int_{b_l}^{b_{l+1}} dbp(K, b \mid D), \tag{A.8}$$

$$p(K, b_l \mid D) = \frac{\exp[-F_n(K, b_l)]}{\sum_{K=0}^{5} \int_{b_1}^{b_L} db \exp[-F_n(K, b)]}, \tag{A.9}$$

$$P(K \mid D) = \sum_{l=1}^{L-1} P(K, b_l \leq b \leq b_{l+1} \mid D), \tag{A.10}$$

$$P(b_l \leq b \leq b_{l+1} \mid D) = \int_{b_l}^{b_{l+1}} dbp(b \mid D), \tag{A.11}$$

$$p(b_l \mid D) = \sum_{K=0}^{5} p(K, b_l \mid D), \tag{A.12}$$

where the integration along the b-axis is calculated using the trapezoidal rule. Note that $\exp[-F_n(K, b_1)] = Z_n(K, b_1) = 0$. The (A) joint probability of (K, b) and the marginal probability of b, (B) the marginal probability of K, and (C) the marginal probability density of b are shown in Fig. A.1. The horizontal axes represent b on a log scale. The colored stairstep graphs and the black one in (A) respectively show the joint probability $P(K, b_l \leq b \leq b_{l+1} \mid D)$ for each K and the marginal probability $P(b_l \leq b \leq b_{l+1} \mid D)$. The three colored graphs of $K < 3$ almost overlap in contrast to Fig. 2.2(A-1). The black bar in (B) shows the marginal probability $P(K \mid D)$. The black markers and black dotted line in (C) respectively show the marginal probability density $p(b_l \mid D)$ and the interpolated values. The vertical black dashed lines and vertical black dash-dotted ones respectively show the true value $b = b_0$ and the estimated value $b = \hat{b}$, as in Fig. 2.2. Both b_0 and $\hat{b}$ are within the same interval of b, which maximize the probabilities $P(K, b_l \leq b \leq b_{l+1} \mid D)$ and $P(b_l \leq b \leq b_{l+1} \mid D)$ in this case. Although the value of K that maximizes $P(K \mid D)$ is the same as $\hat{K}$ in this case, the value of b that maximizes $p(b \mid D)$ is slightly different from $\hat{b}$ in the strict sense. These values are not always consistent in practice, and there is a continuous discussion: which is better, to optimize or to integrate out? [81] The users of our framework can choose a better way in light of their perspective.

A.3 Interpolation of posterior distribution

The density of states in the i^{th} bin, which is the function $g(E; K)$ given the value of μ_k in the interval $[X_i, X_{i+1}]$, is defined and estimated as

$$g(E; K, X_i \leq \mu_k \leq X_{i+1}) := \int dw' \delta[E - E_n(w'; X_i \leq \mu_k \leq X_{i+1})] \varphi(w' \mid K) \tag{A.13}$$

$$= \frac{\sum_{l=1}^{L} N_l(E; X_i \leq \mu_k \leq X_{i+1})}{\sum_{l'=1}^{L} M_{l'}^{(i)} \tilde{Z}_n(K, b_{l'})^{-1} \exp(-n b_{l'} E)}, \tag{A.14}$$

then we obtain

$$\tilde{z}_n(K, b, X_i \leq \mu_k \leq X_{i+1}) = \int dE g(E; K, X_i \leq \mu_k \leq X_{i+1}) \exp(-nbE) \tag{A.15}$$

$$= \sum_{l=1}^{L} \sum_{m=1}^{M_l^{(i)}} \frac{1}{\sum_{l'=1}^{L} M_{l'}^{(i)} \tilde{z}_n(K, b_{l'}, X_i \leq \mu_k \leq X_{i+1})^{-1} \exp\left[n(b - b_{l'}) E_{l,m}^{(i)}\right]}, \quad \text{(A.16)}$$

where $E_n(w'; X_i \leq \mu_k \leq X_{i+1})$, $N_l(E; X_i \leq \mu_k \leq X_{i+1})$, and $E_{l,m}^{(i)}$ respectively indicate $E_n(w)$, $N_l(E)$, and $E_{l,m}$ in the i^{th} bin. $M_l^{(i)}$ is defined as $M_l^{(i)} := \int dE N_l(E; X_i \leq \mu_k \leq X_{i+1})$, where $M_l = \sum_{i=1}^{n-1} M_l^{(i)}$. The values of $\{\tilde{z}_n(K, b_l, X_i \leq \mu_k \leq X_{i+1})\}_{l=1}^{L}$ for each i are determined self-consistently by iterating Eq. (A.16) with $b = b_l$. Given $\{\tilde{z}_n(K, b_l, X_i \leq \mu_k \leq X_{i+1})\}_{l=1}^{L}$ for each i, we calculate $\tilde{z}_n(K, b, X_i \leq \mu_k \leq X_{i+1})$ for each i with $b = b'$ via Eq. (A.16) again. If Δx is sufficiently small (or $\varphi(\mu_k)$ is almost flat), $P(X_i \leq \mu_k \leq X_{i+1} \mid D, K, b)$ is expressed as

$$P(X_i \leq \mu_k \leq X_{i+1} \mid D, K, b) = \frac{\tilde{z}_n(K, b, X_i \leq \mu_k \leq X_{i+1}) \varphi(\mu_k = X_i)}{\sum_{i=1}^{n} \tilde{z}_n(K, b, X_i \leq \mu_k \leq X_{i+1}) \varphi(\mu_k = X_i)}. \quad \text{(A.17)}$$

Appendix B

Appendix in Chapter 3

第 3 章は雑誌掲載が予定される内容を含むため，インターネット公表できません.

Appendix C

Appendix in Chapter 4

第４章は雑誌掲載が予定される内容を含むため，インターネット公表できません.