IPSJ Transactions on Mathematical Modeling and Its Applications Vol.14 No.3 93–101 (Aug. 2021)

Regular Paper

Bayesian Inference for Mixture of Sparse Linear Regression Model

Tomoya Hirakawa[1] Kyosuke Matsudaira[1] Kenji Nagata[2]
Junya Inoue[3] Manabu Enoki[4] Masato Okada[1,a)]

Received: November 19, 2020, Revised: January 15, 2021,
Accepted: April 23, 2021

Abstract: This paper proposes a Bayesian Inference for mixture of sparse linear regression models with the exchange Monte Carlo method. Mixture of linear regression model is a hybrid machine learning model that simultaneously performs clustering and linear regression. Mixture of sparse linear regression model imposes sparsity on the regression parameters and is expected to be applied to the analysis of real data in the field of materials science. The proposed method calculates the mixture ratio of each cluster, the label of each data point, and the posterior distribution of the sparse regression parameters by Bayesian inference using the exchange Monte Carlo method. Model selection based on the Bayesian free energy determines the appropriate number of mixtures of clusters. Experiments on artificial data confirmed that we obtained an appropriate posterior distribution of the parameters and showed appropriate model selection results. We applied our method to the data on aluminum alloys in materials science, and model selection and parameter estimation were performed by Bayesian inference.

Keywords: Bayesian inference, mixture of sparse linear regression model, the exchange Monte Carlo method

1. Introduction

Mixture of regression model [1] clusters given data and performs regression within each class. This model extracts the structure that exists behind the data without directly dividing the data space. It is considered to be an important model in data-driven science. Among them, the Mixture of sparse regression model is expected to improve the simplification, the clarification, and the interpretability of the model itself by assuming sparseness for regression parameters in each class.

Various methods have been proposed for the inference of Mixture of sparse regression model. Khalili et al. and Städler et al. derived Expectation Maximization (EM) algorithms for obtaining a Maximum *a posteriori* (MAP) solution by regularizing Least absolute shrinkage and selection operator (LASSO) and Smoothly clipped absolute deviation (SCAD) for the regression parameters [2], [3]. Blekas et al. proposed a mixture of sparse polynomial regression model for time series data [4]. This study assumed a Gaussian distribution for each element of the regression coefficients to induced sparseness, and derived an EM algorithm to obtain a MAP solution. Furthermore, in Ref. [5], a mixture of regression model in which each component is a multi-kernel of Relevance vector machine was proposed and an EM algorithm was derived for obtaining a MAP solution. These studies constitute algorithms for obtaining MAP solutions and cannot handle the uncertainty of the obtained parameters. A method based on the Bayesian information criterion (BIC) has been proposed for the estimation of appropriate mixture numbers [5]. However, it is known that asymptotic normality does not hold in a statistical model with hierarchical structure such as a neural network and a mixture model focused in this study, and belongs to a singular statistical model [6], [7]; therefore, it is doubtful whether BIC is appropriate for model selection.

Zhang et al. performed Bayesian inference for Mixture of sparse linear regression models using Gibbs Sampling and implemented model selection using Reverse Jump Markov chain Monte Carlo (RJMCMC) [8]. Lee et al. also performed Bayesian inference for mixture of sparse linear regression models using Gibbs Sampling and model selection based on BIC and Akaike information criterion (AIC) [9]. However, it is doubtful whether AIC is also appropriate for model selection in singular statistical model.

In this study, we propose an implementation of Bayesian inference for a mixture of sparse linear regression model with sampling by the exchange Monte Carlo method. The exchange Monte Carlo method avoids the local optimal solution and allows us to perform global sampling from the posterior distribution of parameters. Furthermore, it is possible to calculate Bayesian free energy based on the obtained sampling series, which enables us to perform model selection of the number of mixtures in a mixture of linear regression model. The proposed method is validated with artificial data and applied to the problem of regression for material design and properties in materials science to show its

[1] Graduate School of Frontier Science, The University of Tokyo, Kashiwa, Chiba 277–8561, Japan
[2] National Institute for Materials Science (NIMS), Tsukuba, Ibaraki 305–0047, Japan
[3] Research Center for Advanced Science and Technology, The University of Tokyo, Meguro, Tokyo 153–8904, Japan
[4] Graduate School of Engineering, The University of Tokyo, Bunkyo, Tokyo 113–8656, Japan
[a)] okada@edu.k.u-tokyo.ac.jp

© 2021 Information Processing Society of Japan

effectiveness.

The rest of this paper is organized as follows. Section 2 deals with mixture of regression models and mixture of sparse linear regression models with sparsity introduced into the regression coefficients. Furthermore, an exchange Monte Carlo method is described. In Section 3, simulation of the proposed method on artificial data and simulation of the proposed method on experimental data in materials science are described and the results are discussed. Finally, in Section 4, the future prospects are discussed.

2. Bayesian Inference for Mixture of Sparse Linear Regression Model

In this section, we describe the formulation of generative model and Bayesian inference of mixture of sparse linear regression models.

2.1 Mixture of Linear Regression Model

First, we describe the mixture of linear regression model. The model assumes that given data are generated from multiple probability models, and deals with the task of estimating the probability model from which each data point is generated (clustering) as well as the task of estimating the parameters of each probability model. Now suppose that N pairs of d_x-dimensional input vectors $\boldsymbol{x}_n$ and d_y-dimensional output vectors $\boldsymbol{y}_n$ are given ($n = 1, \cdots, N$). Denote this as $\mathcal{D} = \{X, Y\} = \{\boldsymbol{x}_n, \boldsymbol{y}_n\}_{n=1}^N$. Assuming that each data is generated from a mixture of K probability models, let $\boldsymbol{s}_n$ be a K-dimensional discrete variable vector $\boldsymbol{s}_n \in \{0,1\}^K$, such that the element indicating the mixture to which the data belong is 1 and the others are 0. Then, the overall probability model can be obtained using the mixture ratio $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)^T$, $\sum_{k=1}^K \pi_k = 1$ as follows:

$$p(Y, S|X, K, \Theta) = \prod_{n=1}^{N} \prod_{k=1}^{K} \{\pi_k p(\boldsymbol{y}_n|\boldsymbol{x}_n, \theta_k)\}^{s_{nk}}, \tag{1}$$

where θ_k is a parameter of the k-th class, $\Theta = \{\theta_1, \cdots, \theta_K\}$ is a set of parameters, and s_{nk} is the k-th element of the n-th hidden variable $\boldsymbol{s}_n$. The hidden variable $S = \{\boldsymbol{s}_n\}_{n=1}^N$ is assumed to occur stochastically according to the mixture ratio $\boldsymbol{\pi}$. We also assume a linear model $\boldsymbol{y} = W_k\boldsymbol{x}$ between input and output. $W_k \in \mathbb{R}^{d_y \times d_x}$ is a weight parameter of the k-th model, and collectively $W = \{W_k\}_{k=1}^K$. In this model, let $\Theta = \{\boldsymbol{\pi}, W\}$ and estimate the parameters Θ and the hidden variable S.

A graphical model of mixture of linear regression model is shown in **Fig. 1**. From Fig. 1, the data generation process can be written as follows:

(1) Mixture number K is derived from the prior $p(K)$.
(2) Mixture ratio $\boldsymbol{\pi}$ is derived from the prior $p(\boldsymbol{\pi}|K)$.
(3) Weight parameters W are generated from the prior $p(W|K)$.
(4) Input X is given.
(5) Hidden variables S are generated from the prior $p(S|\boldsymbol{\pi})$.
(6) The output Y is generated from the following relation:

$$\boldsymbol{y}_n = \sum_{k=1}^{K} s_{nk} (W_k\boldsymbol{x}_n) + \boldsymbol{\epsilon}_n, \tag{2}$$

where $\boldsymbol{\epsilon}_n$ is a noise vector.

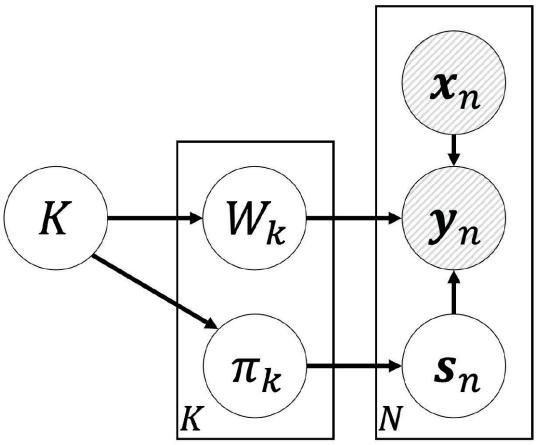

Fig. 1 Graphical model of mixture of linear regression model.

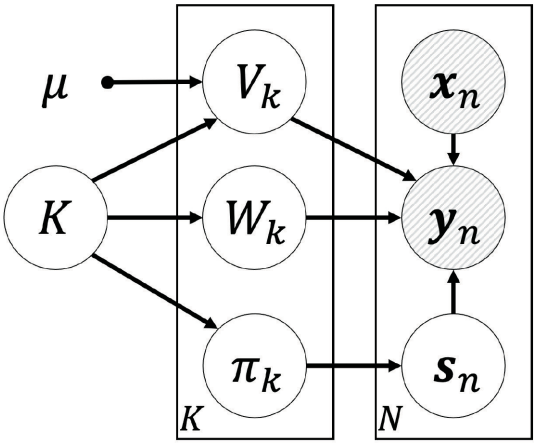

Fig. 2 Graphical model of mixture of sparse linear regression model.

2.2 Introducing Sparsity

This section describes a mixture of sparse linear regression model in which sparsity is imposed on each model parameter of the mixture of linear regression model. We introduce indicator vectors $V = \{V_k\}_{k=1}^K$, $V_k \in \{0,1\}^{d_y \times d_x}$, where each element is a binary variable for each element of the weight parameter in each stochastic model, taking 1 when the element is used and 0 when it is not used. Using the indicator vectors, we define the relationship between each input and output as

$$\boldsymbol{y}_n = \sum_{k=1}^{K} s_{nk} \{(W_k \circ V_k)\boldsymbol{x}_n\} + \boldsymbol{\epsilon}_n, \tag{3}$$

where $\circ$ is the element-wise product. Indicator vectors V are assumed to be generated from the prior $p(V|\mu, K)$ conditioned on the hyper-parameter $\mu \in (0,1)$, which controls the sparseness. A graphical model of this model is shown in **Fig. 2**. As in the previous section, the process of generating the data is

(1) Mixture number K is derived from the prior $p(K)$.
(2) Mixture ratio $\boldsymbol{\pi}$ is derived from the prior $p(\boldsymbol{\pi}|K)$.
(3) Weight parameters W are generated from the prior $p(W|K)$.
(4) Indicator vectors V are generated from the prior $p(V|\mu, K)$.
(5) Input X is given.
(6) Hidden variables S are generated from the prior $p(S|\boldsymbol{\pi})$.
(7) The output Y is generated from the relation (3).

In this model, let $\Theta = \{W, V, \boldsymbol{\pi}\}$ and we estimate Θ and S.

© 2021 Information Processing Society of Japan

2.3 Bayesian Inference

For each input and output $(\boldsymbol{x}_n, \boldsymbol{y}_n)$ in a mixture of sparse linear regression model, we assume that Gaussian noise $\boldsymbol{\epsilon}_n$ with mean $\mathbf{0}$ and variance covariance matrix $\Sigma = \mathrm{Diag}(\sigma_1^2, \cdots, \sigma_{d_y}^2)$ is added. The likelihood function $p(Y, S|X, \Theta, K)$ for the output Y and the hidden variables S in this model is

$$\begin{aligned} p(Y, S|X, \Theta, K) &= \prod_{n=1}^{N} \prod_{k=1}^{K} \{\pi_k p(\boldsymbol{y}_n|\boldsymbol{x}_n, \theta_k)\}^{s_{nk}} \\ &= \prod_{n=1}^{N} \prod_{k=1}^{K} \{\pi_k \mathcal{N}(\boldsymbol{y}_n|(W_k \circ V_k)\boldsymbol{x}_n, \Sigma)\}^{s_{nk}}, \end{aligned} \tag{4}$$

where $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ is the Gaussian distribution of the mean μ and variance covariance matrix Σ. Then, the error function $E(\Theta, S)$ of this model is defined by negative logarithm of likelihood function $p(Y, S|X, \Theta, K)$ as follows,

$$\begin{aligned} E(\Theta, S; K) = &\sum_{n=1}^{N} \sum_{k=1}^{K} s_{nk} \sum_{i=1}^{d_y} \frac{1}{2\sigma_i^2} \left\{y_{ni} - (\boldsymbol{w}_{ik} \circ \boldsymbol{v}_{ik})^T \boldsymbol{x}_n\right\}^2 \\ &+ \frac{N}{2} \sum_{i=1}^{d_y} \log 2\pi\sigma_i^2 - \sum_{k=1}^{K} N_k \log \pi_k, \end{aligned} \tag{5}$$

where N_k is the number of data points belonging to k-th class, $\boldsymbol{w}_{ik}$ and $\boldsymbol{v}_{ik}$ are i-th row of W_k and V_k respectively, and y_{ni} is i-th element of $\boldsymbol{y}_n$. In Bayesian inference, each parameter is treated as a random variable. Firstly, we assume that the value of K is given. The posterior distribution $p(\Theta, S|\mathcal{D}, K)$ of the parameters $\Theta = \{W, V, \boldsymbol{\pi}\}$ and the hidden variable S given the training data $\mathcal{D} = \{X, Y\}$ and the number K of mixture can be written using Bayesian theorem as follows:

$$\begin{aligned} p(\Theta, S|\mathcal{D}, K) &= \frac{p(Y, S|X, \Theta, K)p(\Theta|K)}{p(Y|X, K)} \\ &= \frac{1}{Z_K(\mathcal{D})} \exp\left(-E(\Theta, S; K)\right) p(\Theta|K). \end{aligned} \tag{6}$$

where $p(\Theta|K)$ represents the prior distribution of the parameter Θ and the normalization constant $Z_K(\mathcal{D})$, also called the marginal likelihood, is expressed in the following way;

$$Z_K(\mathcal{D}) = \int \exp\left(-E(\Theta, S; K)\right) p(\Theta|K) \mathrm{d}\Theta \mathrm{d}S. \tag{7}$$

In Bayesian inference, the negative logarithmic marginal likelihood $-\log Z_K(\mathcal{D})$ is called Bayesian free energy, and it is used in model selection. From the Bayesian free energy, posterior distribution $p(K|\mathcal{D})$ of the number of mixture can be calculated as follows:

$$\begin{aligned} p(K|\mathcal{D}) &= \frac{p(\mathcal{D}|K)p(K)}{p(\mathcal{D})} \\ &\propto p(\mathcal{D}|K)p(K) = Z_K(\mathcal{D})p(K), \end{aligned} \tag{8}$$

where $p(K)$ is the prior distribution of the number of mixture K. However, it is difficult to analytically calculate the Bayesian free energy because of the integration of parameters. In this study, Bayesian free energy is calculated numerically using the exchange Monte Carlo method.

2.4 Calculation of Bayesian Free Energy

The Markov chain Monte Carlo method enables us to obtain the normalized constant such as the marginal likelihood in Eq. (7). An auxiliary variable β is introduced and $z_K(\beta)$ is defined as follows:

$$z_K(\beta) = \int \exp\left(-\beta E(\Theta, S; K)\right) p(\Theta|K) \mathrm{d}\Theta \mathrm{d}S. \tag{9}$$

Here, $0 \leq \beta \leq 1$ is a parameter called inverse temperature, and from the definition, $z_K(0) = 1$, $z_K(1) = Z_K(\mathcal{D})$ is clear. To obtain $z_K(1)$ numerically, we consider the inverse temperature sequence $0 = \beta_1 < \beta_2 < \cdots < \beta_{L-1} < \beta_L = 1$, and

$$\begin{aligned} z_K(1) &= \frac{z_K(\beta_L)}{z_K(\beta_{L-1})} \times \frac{z_K(\beta_{L-1})}{z_K(\beta_{L-2})} \times \cdots \times \frac{z_K(\beta_2)}{z_K(\beta_1)} \\ &= \prod_{l=1}^{L-1} \frac{z_K(\beta_{l+1})}{z_K(\beta_l)} \\ &= \prod_{l=1}^{L-1} \frac{\int \exp\left(-\beta_{l+1} E(\Theta, S; K)\right) p(\Theta|K) \mathrm{d}\Theta \mathrm{d}S}{\int \exp\left(-\beta_l E(\Theta, S; K)\right) p(\Theta|K) \mathrm{d}\Theta \mathrm{d}S} \\ &= \prod_{l=1}^{L-1} \left\langle \exp\left(-(\beta_{l+1} - \beta_l) E(\Theta, S; K)\right) \right\rangle_{q(\theta, S; \beta_l)}. \end{aligned} \tag{10}$$

This indicates that the marginal likelihood is obtained as the expected value of the following probability distribution:

$$q(\Theta, S; \beta) \propto \exp\left(-\beta E(\Theta, S; K)\right) p(\Theta|K). \tag{11}$$

Using the exchange Monte Carlo method, it is possible to obtain the value of Eq. (10) depending on the samples obtained [10].

2.5 The Exchange Monte Carlo Method

The exchange Monte Carlo method is one of the Markov chain Monte Carlo methods, which enables us to sample around the global optimal solution even in problems with local solutions. The specific algorithm of the exchange Monte Carlo method is shown below.

(1) We perform Monte Carlo sampling such as the conventional metropolis method or Gibbs sampling from multiple probability distributions $\{q(\Theta_l, S_l; \beta_l)\}_{l=1}^{L}$.

(2) Decide probabilistically whether or not to exchange the parameters $\{\Theta_l, S_l\}$, $\{\Theta_{l+1}, S_{l+1}\}$, of the neighboring distributions with the following probability u.

$$\begin{aligned} u &= \min(1, v) \\ v &= \frac{q(\Theta_{l+1}, S_{l+1}; \beta_l) q(\Theta_l, S_l; \beta_{l+1})}{q(\Theta_l, S_l; \beta_l) q(\Theta_{l+1}, S_{l+1}; \beta_{l+1})} \\ &= \exp\left((\beta_{l+1} - \beta_l)(E(\Theta_{l+1}, S_{l+1}; K) - E(\Theta_l, S_l; K))\right). \end{aligned}$$

By alternately repeating the procedure (1) and (2) above, a sequence of samples $\{\{\Theta_1, S_1\}, \cdots, \{\Theta_L, S_L\}\}$ at each temperature is obtained.

Label-switching [11], in which the uniqueness of the order of classes is lost by the exchange operation, is a problem in parameter estimation of mixture models using the exchange Monte Carlo method. In this study, we re-labeled with the method shown in Section A.2

© 2021 Information Processing Society of Japan

3. Simulations

In this study, numerical simulations of the proposed method are conducted on artificial data and real data of material science. In this section, the simulations and the results are discussed.

3.1 Numerical Simulation on Artificial Data

Firstly, numerical simulations were conducted on artificial data. For the generation of the artificial data, the number of mixtures was set to $K = 3$, and the dimensions of the input space and output space were set to $d_x = 16$ and $d_y = 1$. The number of training data is set to $N = 300$. The variance of the observed noise is set to $\sigma_1^2 = 0.1$, and the hyper-parameter μ for the sparseness is fixed at $\mu = 0.5$. The prior distribution of each parameter is set as follows:

$$p(\Theta|K) = p(W|K)p(V|K,\mu)p(\boldsymbol{\pi}|K)$$

$$p(W|K) = \prod_{k=1}^{K} p(\boldsymbol{w}_{1k}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{w}_{1k}|\boldsymbol{0}, I), \tag{12}$$

$$p(V|\mu, K) = \prod_{k=1}^{K}\prod_{l=1}^{d_x} \mathrm{Bern}(v_{1kl}|1-\mu) \tag{13}$$

$$= \prod_{k=1}^{K}\prod_{l=1}^{d_x} \mu^{(1-v_{1kl})}(1-\mu)^{v_{1kl}}, \tag{14}$$

$$p(\boldsymbol{\pi}|K) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{1}{\beta(\boldsymbol{\alpha})}\prod_{k=1}^{K} \pi_k^{\alpha_k - 1}, \tag{15}$$

$$\boldsymbol{\alpha} = (1, \cdots, 1)^T \in \mathbb{R}^K, \tag{16}$$

where v_{1kl} is l-th element of $\boldsymbol{v}_{1k}$. The prior distribution $p(K)$ of the number of mixture K is set to the uniform distribution from $K = 1$ to $K = 6$.

Figure 3 shows the artificial data generated according to the graphical model shown in Fig. 2. One hundred simulations were conducted to generate artificial data, perform parameter estimation and model selection. The number of replicas in the exchange Monte Carlo method was set to $L = 96$. 20,000 samples were obtained by the exchange Monte Carlo method, and the first 10,000 samples were discarded as burn-in period. See Section A.1 for the specific updating rules of the exchange Monte Carlo method.

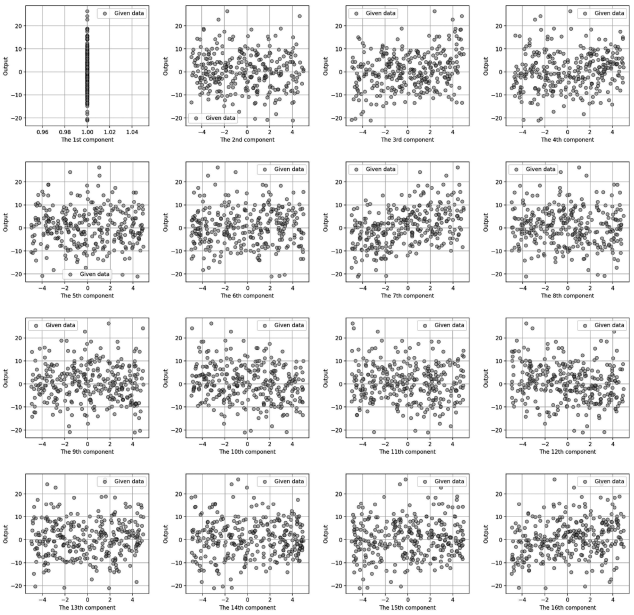

Fig. 3 Examples of generated artificial data ($N = 300$): The horizontal axis of each figure shows the one-dimensional x_i ($i = 1, \cdots, d_x$) with input variables, and the vertical axis shows the output y. The first element of the input variables is always taken to be 1. This means that the first element of the weight parameter $\boldsymbol{w}_{1k}$ of each class corresponds to the intercept of the regression hyperplane.

The results of the numerical simulations for artificial data are described below. **Figure 4** shows the results of the model selection based on Bayesian free energy in one case of 100 simulations. From Fig. 4, we can see that the appropriate mixture number $K = 3$ is selected. Therefore, the simulation results for $K = 3$ are presented below. **Figure 5** and **Fig. 6** respectively show the sampling results for the mixture ratio $\boldsymbol{\pi}$ and the weight parameter $W \circ V$. Figure 5 shows that the sampling of the mixture ratio $\boldsymbol{\pi}$ is performed near the true value. In addition, Fig. 6 shows that the sampling results for the weight parameters are around the true value. The posterior distribution of the mixture ratio $\boldsymbol{\pi}$ and the weight parameter $W \circ V$ are unimodal thanks to the appropriate removal of label switching.

Next, in the same simulation in Fig. 3, 50 sets of indicator vectors with high sampling frequency were extracted, and **Fig. 7** shows the indicator vectors V when they were sorted in the order of frequency. Figure 7 shows that the most sampled indicator vectors match the true indicator vectors. This indicates that the proposed method performs accurate variable selection.

Finally, the results of 100 times model selection experiments are shown in **Fig. 8**. Figure 8 shows that the correct mixture number $K = 3$ could be selected 98 times out of 100 times. This

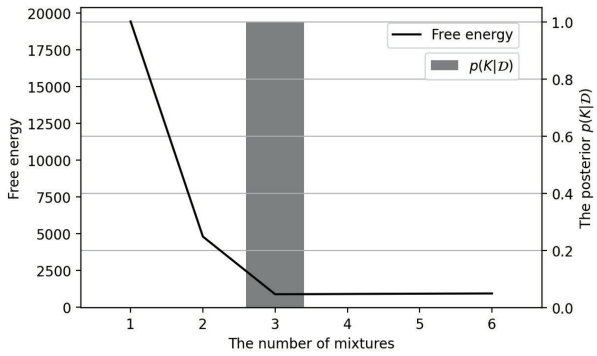

Fig. 4 Bayesian free energy and posterior probability of mixture number K: The line in the figure shows Bayesian free energy, and the bar chart shows the posterior probability of mixture number K, $p(K|\mathcal{D})$.

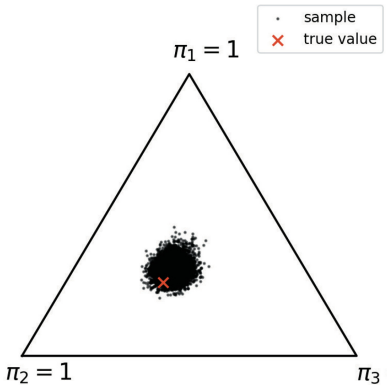

Fig. 5 Sampling results for mixture ratio $\boldsymbol{\pi}$ ($K = 3$).

© 2021 Information Processing Society of Japan

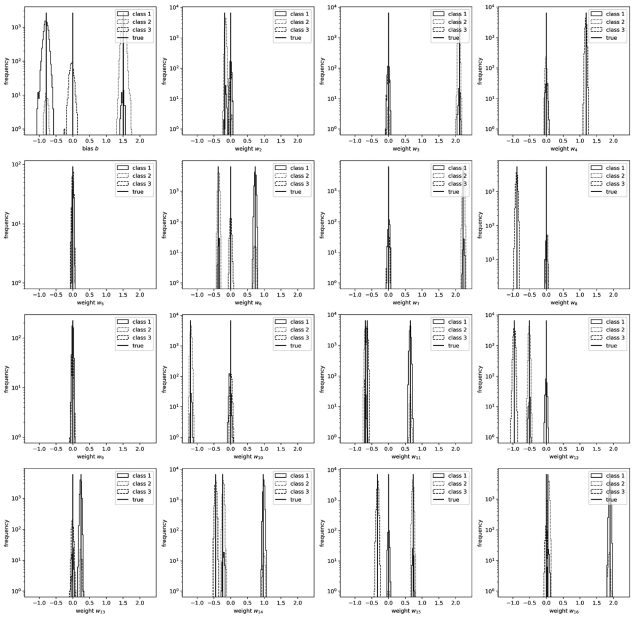

Fig. 6 Sampling results for the weight parameter $W \circ V$ ($K = 3$): The horizontal axis shows the axis w_i ($i = 1, \cdots, d$) for each dimension of the weight parameter and the vertical axis shows the sampling frequency. Values where the corresponding element of indicator vectors is zero are excluded from the figures. The vertical line in each figure shows the true value.

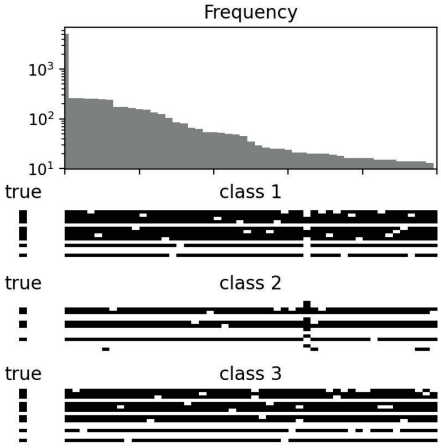

Fig. 7 The pattern of 50 frequently sampled indicator vectors: The top panel shows the number of times sampled, and the three right panels below show the actual indicator vectors sampled, respectively. The three left plots show the true indicator variables, where black color indicates 0 and white color indicates 1.

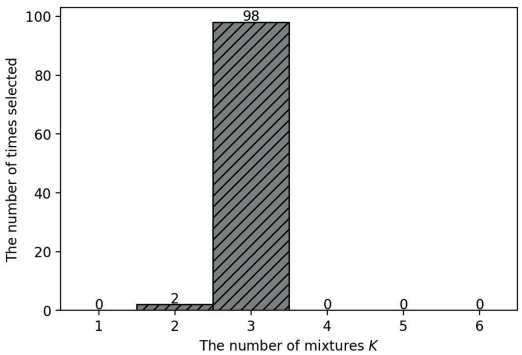

Fig. 8 Results of 100 times model selection simulations: The horizontal axis shows the number of mixtures K, and the vertical axis shows the number of times the mixtures were selected by Bayesian free energy.

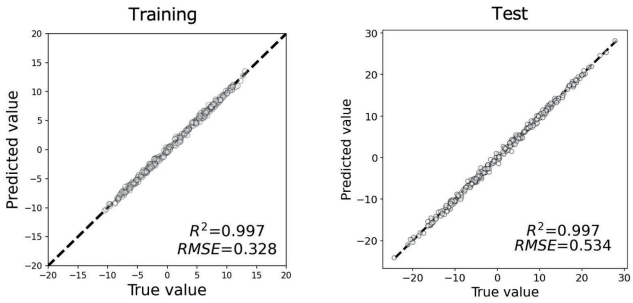

Fig. 9 The regression results of training and test data: the left hand side shows regression result for training data, and the right hand side shows that of test data. The horizontal axis and the vertical of each graph represent the true output and the output predicted from the input data using the trained parameters, respectively.

shows that the proposed method can estimate the correct mixture number from Bayesian free energy.

We conduct additional simulations to confirm prediction performance of our model. Three hundred test input vectors and their class labels are generated in random. In simulations on artificial data set, since the true weight parameters are known, we obtain true output values. Predicted outputs are calculated using optimal weight parameters: the weight parameters are obtained sample series of the exchange Monte Carlo method, and minimize error function for the training data the most. From **Fig. 9**, we confirm our method can correctly predict output value.

3.2 Numerical Simulation for Material Data

In this simulation, the proposed method is applied to the experimental data summarizing the manufacturing conditions and product characteristics of 7,000-series aluminum alloys in materials science. The input is a 16-dimensional variable, $d_x = 16$, that corresponds to the composition and process conditions. The output is a 3-dimensional vector, $d_y = 3$, representing the function of the aluminum alloy. Summarizing our simulation setting, we use 17-dimensional weight parameters $W = \{W_k \in \mathbb{R}^3\}_{k=1}^K$, and indicator vectors $V = \{\boldsymbol{v}_k \in \{0,1\}^{3\times 17}\}_{k=1}^K$, which are added to the intercept parameters. In order to guarantee the confidentiality of the data, each input and output name is discussed without mentioning them, and the experimental values are also discussed in terms of values that have been pre-processed by standardization and other means. The prior distribution of each parameter is set up as follows:

$$p(\Theta|K) = p(W|K)p(V|K,\mu)p(\boldsymbol{\pi}|K)$$

$$p(W|K) = \prod_{k=1}^{K}\prod_{i=1}^{d_y} p(\boldsymbol{w}_{ik}) = \prod_{k=1}^{K}\prod_{i=1}^{d_y} \mathcal{N}(\boldsymbol{w}_{ik}|\boldsymbol{0}, 10\times I),$$

$$p(V|\mu,K) = \prod_{k=1}^{K}\prod_{i=1}^{d_y}\prod_{l=1}^{d_x} \mathrm{Bern}(v_{ikl}|1-\mu)$$

$$= \prod_{k=1}^{K}\prod_{i=1}^{d_y}\prod_{l=1}^{d_x} \mu^{(1-v_{ikl})}(1-\mu)^{v_{ikl}},$$

$$p(\boldsymbol{\pi}|K) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{1}{\beta(\boldsymbol{\alpha})}\prod_{k=1}^{K}\pi_k^{\alpha_k-1},$$

$$\boldsymbol{\alpha} = (1,\cdots,1)^T \in \mathbb{R}^K.$$

The prior distribution $p(K)$ of the number of mixture K is set to

© 2021 Information Processing Society of Japan

the uniform distribution from $K = 1$ to $K = 12$.

In the simulation, $N = 297$ data points were used. Since the output is multidimensional, the noise variance of the first and second element is set as $\sigma_1^2 = \sigma_2^2 = 0.01$, one of the third element is set as $\sigma_3^2 = 0.04$, and the value of the hyper-parameter μ is fixed at $\mu = 0.5$. The number of replicas of the exchange Monte Carlo method was set to $L = 128$. The appropriate number of mixtures is estimated between $K = 1$ to $K = 12$. We obtained 50,000 samples by the exchange Monte Carlo method and discarded the first 25,000 samples as burn-in period.

Simulation results for model selection with free energy are shown in **Fig. 10**. From Fig. 10, we can see that the appropriate mixture number is $K = 3$.

Therefore, we discuss the simulation results for $K = 3$ below. **Figure 11** shows the sampling results for the mixture ratio $\boldsymbol{\pi}$. Compared to the sampling for the artificial data shown in Fig. 5, there is a large variance in the sampling, but we can see that the sampling was performed mainly at certain points. In addition, **Fig. 12** shows the sampling results of the weight parameters for the first output y_1. Comparing these results to those for the artificial data, The histogram has larger variance than that for artificial data. However, some posterior distribution such as the weight parameter w are peaky and has good confidence accuracy. Hence, we can see that the corresponding input has the importance for regressing the output y_1, which becomes a feedback information

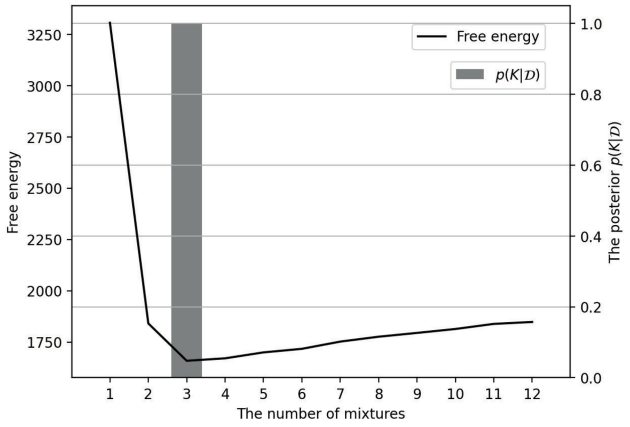

Fig. 10 Bayesian free energy for mixture number K in material data: The line shows the Bayesian free energy, and the bar chart is the posterior probability of the number of mixture K, $p(K|\mathcal{D})$ calculated based on the Bayesian free energy.

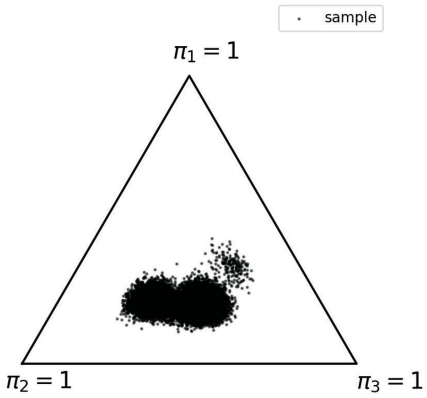

Fig. 11 Sampling results for the mixture ratio $\boldsymbol{\pi}$ at $K = 3$: The black dots in the figure represent the sample.

for experiments for materials science.

Figure 13 shows 50 indicator variables with high sampling frequency, which were extracted and sorted in the order of frequency. Unlike the results for the artificial data in Fig. 7, there was a large variation in the indicator variables sampled, but there was also a large number of specific indicator variables used, suggesting that feature selection by class functioned properly.

Figure 14 illustrates the regression accuracy for the data. The regression performance of the third output is worse than that of the first and second outputs, and this is thought to be due to the fact that a large noise variance was set for the third output only. In order to discuss the appropriateness of this setting, it is important to estimate the noise variance within the framework of Bayesian estimation, and this is an issue to be addressed in the future.

Result of additional simulation is shown in **Fig. 15**. In this simulation, we conduct 10-hold cross validation using the same materials data set. We plot predicted output values of test data of all cross validation subset. From Fig. 15, some of the third outputs have large variance, however, other output values can be predicted roughly.

4. Conclusion

In this paper, we proposed a Bayesian inference for a mixture of sparse linear regression model using the exchange Monte Carlo method. The proposed method is able to obtain appropriate posterior distributions of parameters for artificial data. The number of mixture is appropriately selected by a model selection based on using Bayesian free energy. The proposed method was also applied to the data on aluminum alloys in materials science, and we were able to estimate the appropriate mixture number and each parameter.

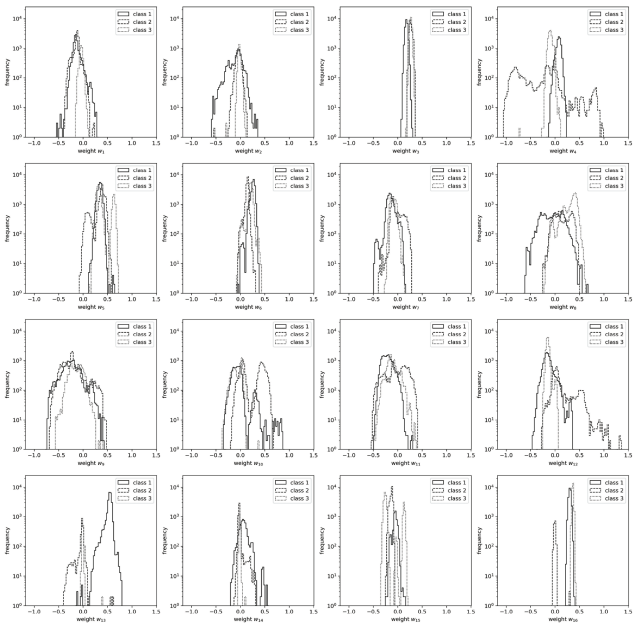

Fig. 12 Sampling results for the weight parameter $W \circ V$ at the first output y_1 at $K = 3$: Histograms of the weight parameters, excluding the intercept parameter, where the horizontal axis of each figure shows the axes of each dimension w_{1i} $(i = 1, \cdots, 16)$ of the weight parameter and the vertical axis shows the frequency of sampling. Samples are excluded from the figures if the corresponding indicator variable is zero.

© 2021 Information Processing Society of Japan

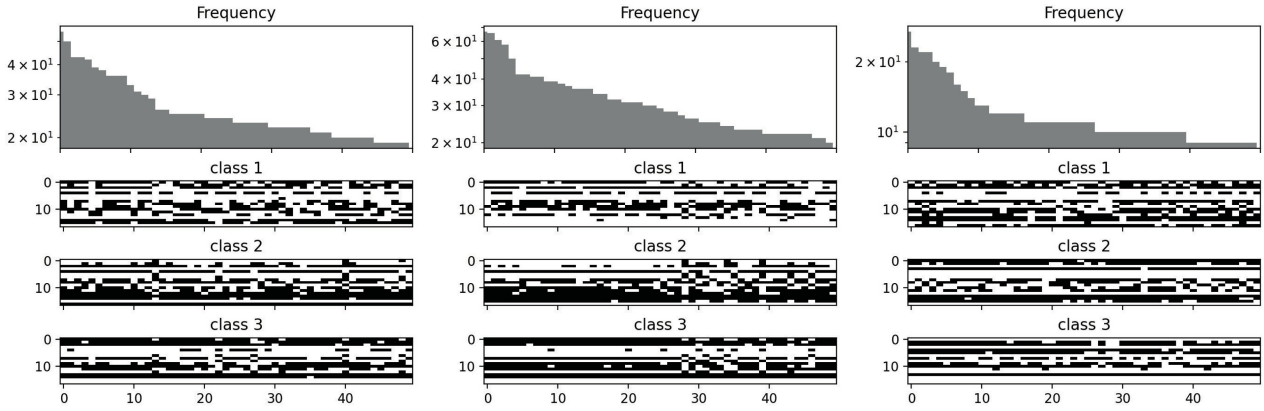

Fig. 13 The 50 most frequently sampled indicator vectors for each output: from left to right, these figures relate to the first, second and third outputs and are viewed in the same way as in Fig. 7.

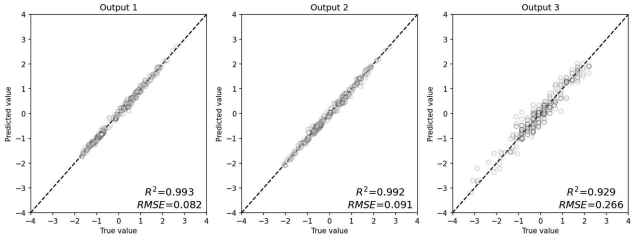

Fig. 14 Graphs showing the fit for the data used: from left to right are the graphs of the first, second and third outputs. The horizontal axis of each graph represents the true output and the vertical axis is the output predicted from the input data using the trained parameters. The closer the line in the figures, the better the fit of the prediction.

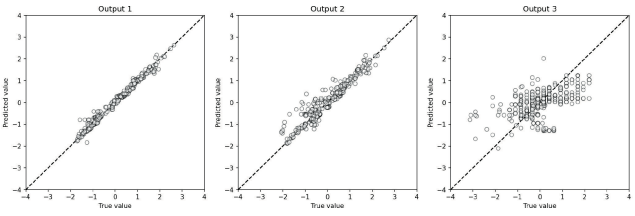

Fig. 15 The result of 10 hold cross validation: from left to right are the graphs of the first, second and third outputs. The horizontal axis and the vertical of each graph represent the true output and the output predicted from the input data using the trained parameters, respectively. The points of all test subsets are shown.

This work constructs an exact Bayesian estimation algorithm of mixture of sparse linear regression model. One of our contributions is that it is possible to compare the effectiveness of previous algorithm such as previous study [5] with that of exact Bayesian estimation. Such comparison is important in constructing faster approximation algorithm, and we address this issue as a future work.

This work deals the mixture of linear regression model, of course we can introduce non-linearity in this model like previous study [9]. In that case, it is important to consider larger framework of model selection problem that discuss appropriate model whether linear or non-linear, which is also addressed as a future work.

In this research, the variances of the noise were treated as a constant. In the future, we need to estimate the noise variance for each class.

Acknowledgments This work was supported by the Council for Science, Technology and Innovation (CSTI), Cross-Ministerial Strategic Innovation Promotion Program (SIP), "Materials Integration for Revolutionary Design System of Structural Materials" (funding agency: JST).

References

[1] Sylvia, R. and Green, P.J.: On Bayesian analysis of mixtures with an unknown number of components (with discussion), *Journal of the Royal Statistical Society*, Series B, Vol.59, No.4, pp.731–792 (1997).
[2] Khalili, A. and Jiahua, C.: Variable selection in finite mixture of regression models, *Journal of the American Statistical Association*, Vol.102, No.479, pp.1025–1038 (2007).
[3] Städler, N., Bühlmann, P. and Van De Geer, S.: ℓ_1-penalization for mixture regression models, *Test*, Vol.19, No.2, pp.209–256 (2010).
[4] Blekas, K., Nikolaos, G. and Aristidis, L.: A sparse regression mixture model for clustering time-series, *Hellenic Conference on Artificial Intelligence*, Springer (2008).
[5] Blekas, K. and Aristidis, L.: Sparse regression mixture modeling with the multi-kernel relevance vector machine, *Knowledge and information Systems*, Vol.39, No.2, pp.241–264 (2014).
[6] Watanabe, S.: Algebraic Analysis for Non-identifiable Learning Machines, *Neural Computation*, Vol.13, No.4, pp.899–933 (2001).
[7] Yamazaki, K. and Watanabe, S.: Singularities in mixture models and upper bounds of stochastic complexity, *International Journal of Neural Networks*, Vol.16, No.7, pp.1029–1038 (2003).
[8] Wei, L. et al.: Model selection in finite mixture of regression models: A Bayesian approach with innovative weighted g priors and reversible jump Markov chain Monte Carlo implementation, *Journal of Statistical Computation and Simulation*, Vol.85, No.12, pp.2456–2478 (2015).
[9] Lee, K.-J., Chen, R.-B. and Wu, Y.N.: Bayesian variable selection for finite mixture model of linear regressions, *Computational Statistics & Data Analysis*, Vol.95, pp.1–16 (2016).
[10] Nagata, K., Sugita, S. and Okada, M.: Bayesian spectral deconvolution with the exchange Monte Carlo method, *Neural Networks*, Vol.28, pp.82–89, Elsevier (2012).
[11] Ajay, J., Holmes, C.C. and Stephens, D.A.: Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling, *Statistical Science*, pp.50–67 (2005).

Appendix

A.1 Sampling Update Rules

In this section, we present the sampling update rules from the probability distribution $q(\Theta;\beta_l)$ of a mixture of sparse linear regression model at each temperature. The graphical model in Fig. 2 shows that the overall distribution of the entire model is

$$\begin{aligned} p(\mathcal{D},\Theta) &= p(X,Y,W,V,S,\boldsymbol{\pi}) \\ &= p(K)p(W|K)p(V|K,\mu)p(\boldsymbol{\pi}|K) \\ &\quad \cdot p(S|\boldsymbol{\pi},K)p(Y|X,W,V,S)p(X). \end{aligned} \tag{A.1}$$

© 2021 Information Processing Society of Japan

The target distribution at inverse temperature β is

$$q(W, V, \boldsymbol{\pi}, S; \beta) \propto \exp\left(-\beta E(\Theta, S; K)\right) p(\Theta|K), \tag{A.2}$$

corresponding to Eq. (9), but it is difficult to sample all the parameters of Θ and S at the same time. Therefore, we perform Gibbs sampling as

$$W^{(i+1)} \sim p(W|V^{(i)}, S^{(i)}, \boldsymbol{\pi}^{(i)}; X, Y), \tag{A.3}$$
$$V^{(i+1)} \sim p(V|W^{(i+1)}, S^{(i)}, \boldsymbol{\pi}^{(i)}; X, Y), \tag{A.4}$$
$$S^{(i+1)} \sim p(S|W^{(i+1)}, V^{(i+1)}, \boldsymbol{\pi}^{(i)}; X, Y), \tag{A.5}$$
$$\boldsymbol{\pi}^{(i+1)} \sim p(\boldsymbol{\pi}|W^{(i+1)}V^{(i+1)}, S^{(i+1)}; X, Y), \tag{A.6}$$

where (i) and $(i+1)$ denote the i, $i+1$-th sample, respectively. In the following subsections, we will show the updating rules for the sampling of Eqs. (A.3) to (A.6).

A.1.1 Update Rules of Weight Parameters W

For the weight parameters W, we perform the following Metropolis update. When only those dependent on the weight parameters W in the overall distribution in Eq. (A.1) are extracted, we get

$$\begin{aligned} &p(W|V^{(i)}, S^{(i)}, \boldsymbol{\pi}^{(i)}; X, Y) \\ &\quad \propto p(Y|X, W, V^{(i)}, S^{(i)})^{\beta} p(W) \\ &\quad = \prod_{n=1}^{N}\prod_{k=1}^{K}\prod_{i=1}^{d_y}\left\{\mathcal{N}\left(y_{ni}|(\boldsymbol{w}_{ik}\circ\boldsymbol{v}_{ik})^T\boldsymbol{x}_n, \sigma_i^2\right)\right\}^{\beta s_{nk}} \\ &\qquad \times \prod_{k=1}^{K}\prod_{i=1}^{d_y}\mathcal{N}(\boldsymbol{w}_{ik}|\boldsymbol{\mu}, \Sigma). \end{aligned} \tag{A.7}$$

Here, the adoption rate of the transitions to the candidate of the i-th sample $W^{(i)}$ with small variations on the i-th sample $W^{(i)}$,

$$W^{(i+1)} = W^{(i)} + \Delta, \tag{A.8}$$

is set to $\min\{1, r\}$ using the

$$r = \frac{p(Y|X, W^{(i+1)}, V^{(i)}, S^{(i)})^{\beta} p(W^{(i+1)})}{p(Y|X, W^{(i)}, V^{(i)}, S^{(i)})^{\beta} p(W^{(i)})}, \tag{A.9}$$

and the adoption of the transitions is determined probabilistically.

A.1.2 Update Rules of Indicator Vectors V

Extracting the terms depending on the indicator vectors V in Eq. (A.1), we get

$$\begin{aligned} &p(V|W^{(i+1)}, S^{(i)}, \boldsymbol{\pi}^{(i)}; X, Y) \\ &\quad \propto p(Y|X, W^{(i+1)}, V, S^{(i)})^{\beta} p(V|\mu) \\ &\quad = \prod_{n=1}^{N}\prod_{k=1}^{K}\prod_{i=1}^{d_y}\left\{\mathcal{N}\left(y_{ni}|(\boldsymbol{w}_{ik}\circ\boldsymbol{v}_{ik})^T\boldsymbol{x}_n, \sigma_i^2\right)\right\}^{\beta s_{nk}} \\ &\qquad \times \prod_{k=1}^{K}\prod_{i=1}^{d_y}\prod_{l=1}^{d_x}\mu^{1-v_{ikl}}(1-\mu)^{v_{ikl}}, \end{aligned} \tag{A.10}$$

where v_{ikl} denotes the l-th element of indicator variables $\boldsymbol{v}_{ik}$. In updating the indicator vectors V, after calculating the posterior probability of each element becoming 0 and 1, respectively, its value is determined according to the posterior probability. Let v_{ikl} be the element of V of interest, the posterior probability for v_{ikl} is

$$\begin{aligned} &p(v_{ikl} = 0|\boldsymbol{w}_{ik}^{(i+1)}, S^{(i)}) \\ &\quad \propto \mu \prod_{k=1}^{K}\left\{\mathcal{N}\left(y_{ni}|(\boldsymbol{w}_{ik}\circ\boldsymbol{v}'_{ik})^T\boldsymbol{x}_n, \sigma_i^2\right)\right\}^{\beta s_{nk}}, \text{ and} \end{aligned} \tag{A.11}$$

$$\begin{aligned} &p(v_{ikl} = 1|\boldsymbol{w}_{ik}^{(i+1)}, S^{(i)}) \\ &\quad \propto (1-\mu) \prod_{k=1}^{K}\left\{\mathcal{N}\left(y_{ni}|(\boldsymbol{w}_{ik}\circ\boldsymbol{v}'_{ik})^T\boldsymbol{x}_n, \sigma_i^2\right)\right\}^{\beta s_{nk}}, \end{aligned} \tag{A.12}$$

respectively. In Eqs. (A.11) and (A.12), $\boldsymbol{v}'_{ik}$ is the value of v_{ikl} in $V^{(i)}$ as 0 and 1, respectively. The posterior probability $p(v_{ikl}|W^{(i+1)}, S^{(i)}, \boldsymbol{\pi}^{(i)}; X, Y)$ can be obtained by calculating the values of Eqs. (A.11) and (A.12) and normalizing them. According to this posterior probability, the value of $v_{ikl}^{(i+1)}$ is determined probabilistically.

A.1.3 Update Rules of Hidden Variables S

For all the hidden variables $\boldsymbol{s}_n$ $(n = 1, \cdots, N)$ corresponding to all data points, the posterior probabilities are calculated and probabilistically updated using the following procedure. In Eqs. (A.1), the terms depending on the hidden variable $\boldsymbol{s}_n$ is extracted as follows:

$$\begin{aligned} &p(\boldsymbol{s}_n|W^{(i+1)}, V^{(i+1)}, \boldsymbol{\pi}^{(i)}; X, Y) \\ &\quad \propto \prod_{k=1}^{K}\prod_{i=1}^{d_y}\left\{p(y_{ni}|\boldsymbol{x}_n, \boldsymbol{w}_{ik}^{(i+1)}, \boldsymbol{v}_{ik}^{(i+1)})\right\}^{\beta s_{nk}} p(\boldsymbol{s}_n|\boldsymbol{\pi})^{\beta} \\ &\quad = \prod_{k=1}^{K}\prod_{i=1}^{d_y}\pi^{\beta s_{nk}}\mathcal{N}(y_{ni}|(\boldsymbol{w}_{ik}^{(i+1)}\circ\boldsymbol{v}_{ik}^{(i+1)})^T\boldsymbol{x}_n, \sigma_i^2)^{\beta s_{nk}}. \end{aligned} \tag{A.13}$$

Calculate this value for all possible $\boldsymbol{s}_n$ values, and by normalizing the sum to be 1, we obtain the posterior probability $p(\boldsymbol{s}_n|W^{(i+1)}, V^{(i+1)}, \boldsymbol{\pi}^{(i)}; X, Y)$ of $\boldsymbol{s}_n$. Using this posterior probability, $\boldsymbol{s}_n$ is determined probabilistically.

A.1.4 Update Rules of Mixture Ratio $\boldsymbol{\pi}$

The mixture ratio $\boldsymbol{\pi}$ is assumed to be a Dirichlet distribution as a prior distribution. Therefore, the posterior distribution of the mixture ratio $\boldsymbol{\pi}$ in simulations of our study is also the following Dirichlet distribution:

$$\begin{aligned} &p(\boldsymbol{\pi}|W^{(i+1)}V^{(i+1)}, S^{(i+1)}; X, Y) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}), \\ &\boldsymbol{\alpha} = (\beta N_1 + 1, \beta N_2 + 1, \cdots, \beta N_K + 1)^T, \end{aligned} \tag{A.14}$$

where N_k is the number of data points belonging to class k. We sample from the above distribution when updating $\boldsymbol{\pi}$.

A.2 Removal of Label Switching

The exchange of parameters between neighboring temperatures in the exchange Monte Carlo method results in uncertainty in the class label order in the samples. In this study, we sort the class labels to achieve consistency in the class label order between samples. We now briefly explain the case where $K = 3$. Let $\Theta^* = \{S^*, W^*, V^*, \boldsymbol{\pi}^*\}$ be the parameter sample that best minimizes the errors defined in Eq. (5). With Θ^* as the reference point, we perform the following sorting:

© 2021 Information Processing Society of Japan

(1) For $\boldsymbol{w}^*_{i1} \circ \boldsymbol{v}^*_{i1}$ and $\boldsymbol{w}^i_{ik} \circ \boldsymbol{v}^i_{ik}$ in the i-th sample, we assign label 1 to the class k_1 that maximizes its inner product,

(2) Similarly, for two classes other than class k_1, we compute inner product of $(\boldsymbol{w}^*_{i2} \circ \boldsymbol{v}^*_{i2})$ and assign label 2 to the larger class and label 3 to the smaller one.

$$k_1 = \arg\max_k (\boldsymbol{w}^*_{i1} \circ \boldsymbol{v}^*_{i1})^T (\boldsymbol{w}^i_{ik} \circ \boldsymbol{v}^i_{ik}) \tag{A.15}$$

(3) Repeat (1) and (2) above for all samples.

By sorting in this way, we can solve the uncertainty of class label order in the exchange Monte Carlo method.

Tomoya Hirakawa was born in 1997. He graduated from National Institute of Technology, Kurume College (NITKC) in 2020. He is currently a M.S. student in Graduate School of Frontier Science from The University of Tokyo. His current research interest is in data driven science with Bayesian Estimation.

Kyosuke Matsudaira received his B.Sc. degree in physics from Tokyo University of Science, Japan, in 2018, and his M.Sc. degree in Graduate School of Frontier Sciences from The University of Tokyo, Japan, in 2020. He is currently a Ph.D. student in Graduate School of Frontier Sciences from The University of Tokyo, where he is developing data analysis methods for inelastic neutron scattering. His current research interests include statistical data analysis, machine learning, condensed matter physics.

Kenji Nagata is currently a senior researcher in National Institute for Materials Science (NIMS), Japan. He received his master and doctoral degree on computer science from Tokyo Institute of Technology in 2006 and in 2008, respectively. His research interest is data-driven science for materials informatics.

Junya Inoue was educated at the University of Tokyo and was Research Associate at Brown University before joining the University of Tokyo as an assistant professor in 2000. He is currently a professor at the Institute of Industrial Science, the University of Tokyo.

Manabu Enoki is Professor of the Department of Materials Engineering at The University of Tokyo. His research interest concerns the analysis of mechanical properties in various materials using physical based model and data-driven approach.

Masato Okada is a professor in Graduate School of Frontier Sciences, The University of Tokyo, Japan. He received his M.Sc. and Ph.D. degrees from Osaka University, Japan, in 1987 and 1997, respectively. From 1987 to 1989, he worked at Mitsubishi Electric Corporation. From 1991 to 1996 he was a research associate at Osaka University. He was a researcher in the Kawato Dynamic Brain Project until 2001. He was a deputy laboratory head in RIKEN Brain Science Institute, Japan, until 2004. His research interests include computational aspects of neural networks and statistical mechanics for information processing.

© 2021 Information Processing Society of Japan