Cluster's Number Free Bayes Prediction of General Framework on Mixture of Regression Models

Haruka Murayama¹ · Shota Saito¹ · Yuji Iikubo¹ · Yuta Nakahara² · Toshiyasu Matsushima¹

Received: 19 January 2021 / Accepted: 31 August 2021 / Published online: 20 September 2021
© The Author(s) 2021

Abstract

Prediction based on a single linear regression model is one of the most common way in various field of studies. It enables us to understand the structure of data, but might not be suitable to express the data whose structure is complex. To express the structure of data more accurately, we make assumption that the data can be divided in clusters, and has a linear regression model in each cluster. In this case, we can assume that each explanatory variable has their own role; explaining the assignment to the clusters, explaining the regression to the target variable, or being both of them. Introducing probabilistic structure to the data generating process, we derive the optimal prediction under Bayes criterion and the algorithm which calculates it sub-optimally with variational inference method. One of the advantages of our algorithm is that it automatically weights the probabilities of being each number of clusters in the process of the algorithm, therefore it solves the concern about selection of the number of clusters. Some experiments are performed on both synthetic and real data to demonstrate the above advantages and to discover some behaviors and tendencies of the algorithm.

Keywords Bayes criterion · Clustering · Linear regression · Variational Bayes algorithm

Abbreviations

Optimal prediction under Bayes criterion	The decision rule $\delta^*(D)$ which minimizes the Bayes risk
BMI	Body mass index

✉ Haruka Murayama
h-murayama@ruri.waseda.jp

Extended author information available on the last page of the article

1 Introduction

In the field of machine learning, prediction problem is a common problem, which uses pairs of explanatory variable and target variable $\{(x_i, y_i)\}_{i=1}^n$ and new explanatory variable x_{n+1} to predict corresponding target variable y_{n+1} . The data analysis using *linear regression model* has been one of the most common and fundamental way to detect the structure of the data. However, it might not be suitable to adopt a linear regression model when the structure of the data is too complex.

To attain high prediction performance with interpretability, previous works have proposed some extensions of linear regression model (e.g., [1–14]). In stratified regression [1], the data are stratified based on design variables, and at each level, a linear regression model is defined using explanatory variables and response variables. Another extension of the linear regression model is Mixture of Experts Model (e.g., [2, 3]) or Hierarchical Mixture of Experts Model (e.g., [4–8]), which consists of some experts and a gating function. Each expert is a linear regression model that outputs the response variable. The gating function, which receives the same input as experts, weights each output of experts, so that we can finally get a single output. Piecewise Linear Regression Model (e.g., [9, 10]) is also an extension of linear regression model. It divides the input space, and has different linear regression model in each subspace. Generalized Linear Mixed Models [11] considers “random effect” in its model, which leads the data to be divided into some groups and have different linear regression model in each group. Although each work has different algorithms, the structure they assume to the data are the same: the data can be grouped into clusters, and have linear regression model in each cluster.

Here, we provide a point of view to organize these previous works. By focusing on the probabilistic model which expresses the data generation, we can treat these previous works in unified manner. Each of the data is represented as $(z, x_1, \dots, x_d, y)$, which is the realized value of random variables $Z, X_1, \dots, X_d$ and Y . Here, $Z \in \{1, \dots, k\}$ is assign variable, which decides the assignment of the data point to k clusters. $X_1, \dots, X_d$ are the explanatory variables, and Y represents the target variable. Here, let U and V be defined as

$$\begin{aligned} U &:= \{X_i \in \{X_1, \dots, X_d\} \mid X_i \text{ explains the cluster structure}\}, \\ V &:= \{X_j \in \{X_1, \dots, X_d\} \mid X_j \text{ explains the regression to } Y\}. \end{aligned}$$

From this point of view, we can consider broad patterns of data generation models by adjusting these factors:

- Number of variables in $U \cap V$.
- Whether Y is continuous or discrete.
- Whether Z is observable or not.

Many of the previous works can be organized by this point of view. For example, the stratified regression is the one with $U = \emptyset$, continuous Y , and observable Z . Mixture of Linear Experts and Piecewise Linear Regression model are the ones with $U \cap V = \{X_1, \dots, X_d\}$, continuous Y , and unobservable Z .

In this paper, we assume the data generation model with continuous target variable Y , and unobservable assignment variable Z . Here, we do not restrict the number of elements in $U \cap V$, therefore we can treat broad patterns of data generation model by our approach. Assuming this structure to the data, we deliver the optimal prediction under Bayes criterion [15]. Moreover, we derive an approximative algorithm that calculates the prediction by adopting variational inference method.

The advantages of using our algorithm are shown below:

1. It is not necessary to input the number of clusters beforehand, since our proposed algorithm automatically weights the probabilities of being each number of clusters in the process of prediction.
2. Since the prediction is based on the explicit assumption of a probabilistic data generation structure, it is easier to distinguish the data structure.
3. We can easily consider broader extensions of our algorithm. For example, by using logistic regression model instead of linear regression model, we can solve classification problem where Y is discrete.

Regarding (1), recent works (e.g., [5–7]) often fix the number of regression models, so the number of clusters. On the other hand, our algorithm do not choose the number of clusters to be a single value, but consider several numbers of clusters and weight their prediction results.

In Sect. 2, we define probabilistic structure which expresses the data generation. Then, in Sects. 3 and 4, we derive the optimal prediction under Bayes criterion and the approximative algorithm respectively. In Sect. 5, we mention some experiments conducted to confirm the behavior of the proposed algorithm using both synthetic and real data. In the first experiment, we synthesized data and parameters from the priors, so that it is possible to check the average behavior of the algorithm. In the second experiment, we generated the data with specific tendency, so that we can confirm more detailed behavior of the algorithm. In the last experiment, we used real data to confirm that our algorithm works well in practical use, too.

2 Probabilistic Model of Data Generation

We assume that the given data have the following characteristics:

- The data are the pairs of explanatory variables and response variables.
- All the variables are continuous.
- It is not appropriate to adopt a single linear regression model to the data.
- If we divide the data into some clusters, it will be appropriate to adopt a linear regression model in each cluster.

Let each of the given data be represented as $(z, \mathbf{x}, y)$, which is the realized value of three random variables $(Z, \mathbf{X}, Y)$. Each random variable has following features.

From now on, we express the p.d.f. of Gaussian distribution, Dirichlet distribution, and Wishart distribution as $\mathcal{N}(\cdot)$, $\text{Dir}(\cdot)$ and $\mathcal{W}(\cdot)$, respectively.

Assumption 2.1 Let K be an unobservable random variable which represents the number of clusters, and k be a realized value of K . Also, let $k_{\max}$ be a fixed value which represents the maximum number of clusters we assume.

Assumption 2.2 Let $\mathbf{Z}_k = (Z_1, \dots, Z_k)^\top \in \{0, 1\}^k$ be a latent random variable called *assignment variable*, given the number of clusters k . Here, the l th component Z_{l_k} ($l_k = 1, \dots, k$) is defined as

$$Z_{l_k} = \begin{cases} 1 & \text{(when in cluster } l_k) \\ 0 & \text{otherwise} \end{cases}, \sum_{l_k=1}^k Z_{l_k} = 1. \quad (1)$$

The distribution of $\mathbf{Z}_k$ is given by

$$p(\mathbf{z}_k \mid \boldsymbol{\pi}_k) = \prod_{l_k=1}^k \pi_{l_k}^{z_{l_k}}, \quad (2)$$

where $\boldsymbol{\pi}_k := (\pi_1, \dots, \pi_k)^\top \in [0, 1]^k$ is a mixing parameter which fulfills

$$\sum_{l_k=1}^k \pi_{l_k} = 1. \quad (3)$$

Assumption 2.3 Let $\mathbf{X} = (X_1, \dots, X_d)^\top \in \mathbb{R}^d$ be an observable random variable, where $d := p + q + r$. Given the number of clusters k , the first $(p + q)$ variables of $\mathbf{X}$ follows the Gaussian mixture distribution as below:

$$p(\mathbf{x} \mid \mathbf{z}_k, \mathbf{M}_k, \mathbf{L}_k) = p(x_1, \dots, x_{p+q})p(x_{p+q+1}, \dots, x_d) \quad (4)$$

$$= \prod_{l_k=1}^k \mathcal{N}(x_1, \dots, x_{p+q} \mid \boldsymbol{\mu}_{l_k}, \boldsymbol{\Lambda}_{l_k}^{-1})^{z_{l_k}} p(x_{p+q+1}, \dots, x_d), \quad (5)$$

where $\mathbf{M}_k := \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k\}$ is a set of $(p + q)$ dimensional mean vectors, and $\mathbf{L}_k := \{\boldsymbol{\Lambda}_1, \dots, \boldsymbol{\Lambda}_k\}$ is a set of $(p + q) \times (p + q)$ dimensional precision matrices.

Equation (5) suggests that among explanatory variables $X_1, \dots, X_d$, the first $(p + q)$ variables follow k mixtures of Gaussian distributions, and the rest r variables follows arbitrary distributions. To give an example in real world data, think of precipitation prediction in some observation areas. The first $p + q$ explanatory variables can be elevation, latitude, and distance from the sea which might explain the cluster structure of observation areas.

Assumption 2.4 Let $Y \in \mathbb{R}$ be an observable random variable called *target variable*. Given the number of clusters k , Y follows the regression model of the last $(q + r)$ explanatory variables :

$$p(y|\mathbf{x}, \mathbf{W}_k, \mathbf{z}_k) = \prod_{l_k=1}^k \mathcal{N}\left(y | w_{l_k,0} + w_{l_k,1}x_{p+1} + \cdots + w_{l_k,(q+r)}x_d, \sigma_{l_k}^2\right)^{z_{l_k}}. \quad (6)$$

Here, $\mathbf{W}_k := (\mathbf{w}_1, \dots, \mathbf{w}_k) \in \mathbb{R}^{(q+r+1) \times k}$ is a coefficient parameter, where

$$\mathbf{w}_{l_k} = (w_{l_k,0}, \dots, w_{l_k,(q+r)})^T \in \mathbb{R}^{q+r+1} \quad (l_k = 1, \dots, k), \quad (7)$$

and $\sigma_{l_k}^2 \in \mathbb{R}_+$ ($l_k = 1, \dots, k$) is a variance which is a fixed value.

Equation (6) suggests that among explanatory variables $X_1, \dots, X_d$, the last $(q + r)$ variables explain the regression to the target variable Y .

As can be understood from Assumptions 2.3 and 2.4, the first $(p + q)$ explanatory variables explain the cluster structure, and the last $(q + r)$ explanatory variables explain the regression to Y . Therefore, we name each variable as below.

Definition 2.1 Regarding Assumptions 2.3 and 2.4, the first $(p + q)$ explanatory variables are called *cluster explanatory variable*, and the last $(q + r)$ explanatory variables are called *regression explanatory variable*, i.e.,

$$\mathbf{X} = (\overbrace{X_1, \dots, X_p, X_{p+1}, \dots, X_{p+q}}^{\text{cluster explanatory variable}}, \underbrace{X_{p+q+1}, \dots, X_{p+q+r}}_{\text{regression explanatory variable}}).$$

Assumption 2.5 Let $\pi_k, \mathbf{M}_k, \mathbf{L}_k, \mathbf{W}_k$, and K have respective priors $p(\pi_k)$, $p(\mathbf{M}_k)$, $p(\mathbf{L}_k)$, $p(\mathbf{W}_k)$, and $p(K)$.

The graphical model of data and parameters given the number of clusters k is shown in Fig. 1.

Definition 2.2 Given the number of clusters k , let parameters θ_k be defined as

$$\theta_k := \{\pi_k, \mathbf{M}_k, \mathbf{L}_k, \mathbf{W}_k\}. \quad (8)$$

Also, let $\mathbf{x}^n$ express n independent, identically distributed data drawn from $p(\mathbf{x})$, i.e., $\mathbf{x}^n := \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, and define $\mathbf{y}^n, \mathbf{z}^n$ in the same way. Then, define given data $\mathbf{D}$ as $\mathbf{D} := \{\mathbf{x}^n, \mathbf{y}^n, \mathbf{x}_{n+1}\}$. Here, new data $(\mathbf{x}_{n+1}, y_{n+1})$ is drawn independently from the identical distribution as $(\mathbf{x}_i, y_i)$ where $i = 1, \dots, n$.

Hereafter, we use calligraphic fonts to express the set where each random variable belongs to.

Fig. 1 The graphical model of data and parameters given the number of clusters k . The white nodes represent the unobserved variables, and the grey nodes represent the observed variables

3 Optimal Prediction Under Bayes Criterion

Assuming the given data to be represented by the probabilistic model introduced in Sect. 2, we propose the optimal prediction in the sense of the Bayes criterion when considering squared-error loss.

Since the goal is to predict the new response variable y_{n+1} from the given data $\mathbf{D}$, the decision rule δ can be defined as follows.

$$\delta : \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{X} \rightarrow \mathcal{Y}; \mathbf{D} \mapsto \delta(\mathbf{D}). \quad (9)$$

Using the squared-error loss between the true value of y_{n+1} and the decision $\delta(\mathbf{D})$, we define the *loss function* L as below:

$$L(\delta(\mathbf{D}), \mathbf{z}_k^{n+1}, \theta_k, k) = \int_{\mathcal{Y}} (y_{n+1} - \delta(\mathbf{D}))^2 p(y_{n+1} | \mathbf{x}_{n+1}, \mathbf{z}_{n+1 k}, \theta_k, k) dy_{n+1}. \quad (10)$$

Given the loss function, the *risk function* of decision rule $\delta(\cdot)$ is defined as follows.

$$R(\delta(\cdot), \mathbf{z}_k^{n+1}, \theta_k, k) := \int_{\mathcal{X}^{n+1}} \int_{\mathcal{Y}^n} L(\delta(\mathbf{D}), \mathbf{D}, \mathbf{z}_k^{n+1}, \theta_k, k) p(\mathbf{D} | \mathbf{z}_k^{n+1}, \theta_k, k) dy^n d\mathbf{x}^{n+1}. \quad (11)$$

Moreover, the *Bayes risk* of decision rule $\delta(\mathbf{D})$, with respect to the prior distributions defined in Assumptions 2.2 and 2.5 is defined as

$$\text{BR}(\delta(\cdot)) := \sum_{k=1}^{k_{\max}} \sum_{\mathcal{Z}_k^{n+1}} \int_{\Theta_k} R(\delta(\mathbf{D}), \mathbf{z}_k^{n+1}, \theta_k, k) p(\mathbf{z}_k^{n+1}, \theta_k, k) d\theta_k. \quad (12)$$

Here, we can obtain the following result. The proof of Proposition 3.1 is in Appendix A.

Proposition 3.1 *The decision rule $\delta^*(\mathbf{D})$ which minimizes the Bayes risk (which we call “optimal prediction under Bayes criterion”) is*

$$\delta^*(\mathbf{D}) = \int_{\mathcal{Y}} y_{n+1} p^*(y_{n+1} | \mathbf{D}) dy_{n+1}, \quad (13)$$

where

$$p^*(y_{n+1} | \mathbf{D}) := \sum_{k=1}^{k_{\max}} \sum_{\mathbf{z}_k^{n+1}} \int_{\Theta_k} p(y_{n+1} | \mathbf{D}, \mathbf{z}_k^{n+1}, \theta_k, k) p(\mathbf{z}_k^{n+1}, \theta_k, k | \mathbf{D}) d\theta_k. \quad (14)$$

Here, $p^*(y_{n+1} | \mathbf{D})$ is called predictive distribution.

Note here, given the same data set $(\mathbf{x}^n, \mathbf{y}^n)$, different new data points like $\mathbf{x}_{n+1}$ and $\mathbf{x}'_{n+1}$ will cause different posterior distributions of the parameters and the latent variables such as k and $\mathbf{z}^{n+1}$. The effect of the new data $\mathbf{x}_{n+1}$ would be smaller when the size of data set n is large.

4 Approximation Using Variational Bayes Algorithm

Although we derived optimal prediction under Bayes criterion in Proposition 3.1, it is hard to obtain a closed-form analytical solution of posterior distribution $p(\mathbf{z}_k^{n+1}, \theta_k, k | \mathbf{D})$ which appears in (14). Therefore, we adopt variational Bayes algorithm to approximate the posterior distribution, and obtain the approximation of optimal prediction under Bayes criterion.

4.1 Priors of Parameters and Cluster Number

First, we introduce priors over each parameter given the number of clusters k ; $\pi_k, \mathbf{M}_k, \mathbf{L}_k, \mathbf{W}_k$, and the cluster number K . Here, we mostly choose conjugate distribution for each prior of the parameters, therefore the analysis would be considerably simplified.

Assumption 4.1 Let uniform distribution govern the cluster number K , i.e.,

$$p(k) = \frac{1}{k_{\max}}. \quad (15)$$

Assumption 4.2 Let mixing parameter $\pi_k \in [0, 1]^k$ be drawn over Dirichlet distribution

$$p(\boldsymbol{\pi}_k) = \text{Dir}(\boldsymbol{\pi}_k \mid \boldsymbol{\alpha}_{0_k}) = C(\boldsymbol{\alpha}_{0_k}) \prod_{l_k=1}^k \pi_{l_k}^{\alpha_{0_k} - 1}, \quad (16)$$

where $\boldsymbol{\alpha}_{0_k} := (\alpha_{0_k}, \dots, \alpha_{0_k})^\top \in \mathbb{R}^k$ is real-valued constant vector, and $C(\boldsymbol{\alpha}_{0_k})$ is the normalization constant on Dirichlet distribution.

Assumption 4.3 Let Gaussian-Wishart distribution govern the mean vector and precision matrix of each Gaussian component, i.e.,

$$p(\mathbf{M}_k, \mathbf{L}_k) = p(\mathbf{M}_k \mid \mathbf{L}_k) p(\mathbf{L}_k) = \prod_{l_k=1}^k \mathcal{N}(\boldsymbol{\mu}_{l_k} \mid \mathbf{m}_{0_k}, (\beta_{0_k} \boldsymbol{\Lambda}_{l_k})^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_{l_k} \mid \mathbf{A}_{0_k}, \nu_{0_k}). \quad (17)$$

Here, all the hyper-parameters $\mathbf{m}_{0_k} \in \mathbb{R}^{p+q}$, $\beta_{0_k} \in \mathbb{R}_+$, $\mathbf{A}_{0_k} \in \mathbb{R}^{(p+q) \times (p+q)}$, and $\nu_{0_k} \in \mathbb{R}$ are constant.

Assumption 4.4 Let coefficient parameter be drawn over Gaussian distribution in each cluster, i.e.,

$$p(\mathbf{W}_k) = \prod_{l_k=1}^k \mathcal{N}(\mathbf{w}_{l_k} \mid \boldsymbol{\mu}_{\mathbf{w}0_k}, \boldsymbol{\Lambda}_{\mathbf{w}0_k}^{-1}). \quad (18)$$

Here, mean vector $\boldsymbol{\mu}_{\mathbf{w}0_k} \in \mathbb{R}^{q+r+1}$, and precision matrix $\boldsymbol{\Lambda}_{\mathbf{w}0_k} \in \mathbb{R}^{(q+r+1) \times (q+r+1)}$ are both constant.

4.2 Analysis on Approximated Posterior Distribution

Now, we consider an *approximated posterior distribution*. Approximated posterior distribution¹ is a distribution of the latent variables $\mathbf{Z}_k^{n+1}$, the parameters $\boldsymbol{\theta}_k$, and the cluster number K , which belongs to the function family $\mathfrak{Q}$ defined as below:

$$\mathfrak{Q} := \{q(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k, k) \mid q(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k, k) = q(\mathbf{z}_k^n) q(\mathbf{z}_{n+1}^k) q(\boldsymbol{\pi}_k) q(\mathbf{M}_k, \mathbf{L}_k) q(\mathbf{W}_k) q(k)\}. \quad (19)$$

The purpose of the variational Bayes algorithm is to obtain the variational distribution $q^* \in \mathfrak{Q}$ which minimizes its Kullback–Leibler (KL) divergence between posterior distribution $p(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k, k \mid \mathbf{D})$, i.e.,

$$q^*(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k, k) = q^*(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k \mid k) q^*(k) \quad (20)$$

¹ It is called “variational distribution” in [16].

$$= \arg \min_{q \in \mathcal{Q}} \sum_{k=1}^{k_{\max}} \sum_{\mathcal{Z}_k^{n+1}} \int_{\Theta_k} q(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k, k) \ln \frac{q(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k, k)}{p(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k, k | \mathbf{D})} d\boldsymbol{\theta}_k. \quad (21)$$

Here, we can obtain the following two propositions.²

$$\ln q^*(\mathbf{z}_k^n) := E_{q^*(\mathbf{z}_k^n)} [\ln p(\mathbf{x}^{n+1}, y^n, \mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k | k)] + \text{const.}, \quad (22)$$

Proposition 4.1

$$\ln q^*(\mathbf{z}_{n+1, k}) := E_{q^*(\mathbf{z}_{n+1, k})} [\ln p(\mathbf{x}^{n+1}, y^n, \mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k | k)] + \text{const.}, \quad (23)$$

$$\ln q^*(\boldsymbol{\pi}_k) := E_{q^*(\boldsymbol{\pi}_k)} [\ln p(\mathbf{x}^{n+1}, y^n, \mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k | k)] + \text{const.}, \quad (24)$$

$$\ln q^*(\mathbf{M}_k, \mathbf{L}_k) := E_{q^*(\mathbf{M}_k, \mathbf{L}_k)} [\ln p(\mathbf{x}^{n+1}, y^n, \mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k | k)] + \text{const.}, \quad (25)$$

$$\ln q^*(\mathbf{W}_k) := E_{q^*(\mathbf{W}_k)} [\ln p(\mathbf{x}^{n+1}, y^n, \mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k | k)] + \text{const.} \quad (26)$$

Then,

$$q^*(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k | k) = q^*(\mathbf{z}_k^n) q^*(\mathbf{z}_{n+1, k}) q^*(\boldsymbol{\pi}_k) q^*(\mathbf{M}_k, \mathbf{L}_k) q^*(\mathbf{W}_k), \quad (27)$$

where $E_{q^*(\star)}[\cdot]$ denotes an expectation with respect to the distribution q^* over all variables except $\star$.

Proposition 4.2 Regarding variational distribution of K , it holds that

$$\ln q^*(k) = \ln p(k) + \ln k! + \sum_{\mathcal{Z}_k^{n+1}} \int_{\Theta_k} q^*(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k | k) \ln \frac{p(\mathbf{x}^{n+1}, y^n, \mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k | k)}{q^*(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k | k)} d\boldsymbol{\theta}_k + \text{const.} \quad (28)$$

Equations (27) and (28) give the distribution which minimizes the KL divergence between the posterior distribution subject to the constraint of $\mathcal{Q}$. However, they still do not give the explicit solution, since the expectations in (22) to (26) depend on the other factors of q^* . Therefore, we first initialize all the factors and repeat updating each factor in turn using (22)–(26) so that we can approximately calculate (27).

In the next subsections, we will introduce updating formula of each factors $q^*(\cdot_k)$. Hereafter, $q^{(t)}(\star_k)$ denotes the variational distribution of latent variables or parameter $\star$ given cluster number k at time t , and $E_{\star_k}[\cdot]$ denotes the expectation with respect to $q^{(t)}(\star_k)$. Also, for simplicity, we will define $\mathbf{U} \in \mathbb{R}^{p+q}$ and $\mathbf{V} \in \mathbb{R}^{q+r}$ as following:

² Refer Section 10.1.1. and 10.1.4 of [16] to see the proofs of Propositions 4.1 and Section 10.1.4 of 4.2, respectively.

$U := (X_1, \dots, X_{p+q})$ (cluster explanatory variables),

$V := (X_{p+1}, \dots, X_d)$ (regression explanatory variables).

We use $\mathbf{u}_i = (u_{i1}, \dots, u_{i(p+q)})^\top \in \mathbb{R}^{p+q}$, $\mathbf{v}_i = (v_{i1}, \dots, v_{i(q+r)})^\top \in \mathbb{R}^{q+r}$ ($i = 1, \dots, n$) as the realized values of U and V respectively, and define $\tilde{\mathbf{v}}_i$ as $\tilde{\mathbf{v}}_i = (1, v_{i1}, \dots, v_{i(q+r)}) \in \mathbb{R}^{q+r+1}$ ($i = 1, \dots, n$).

1. Updating $q(\mathbf{z}_k^n)$: From (22), the updating formula of $q(\mathbf{z}_k^n)$ is given by

$$q^{(t+1)}(\mathbf{z}_k^n) = \prod_{i=1}^n \prod_{l_k=1}^k r_{il_k}^{(t) z_{il_k}}. \quad (29)$$

Here, $r_{il_k}^{(t)}$ is given by

$$r_{il_k}^{(t)} = \frac{\rho_{il_k}^{(t)}}{\sum_{l'_k=1}^k \rho_{il'_k}^{(t)}}, \quad (30)$$

where

$$\begin{aligned} \rho_{il_k}^{(t)} = \exp \left[\frac{1}{2} \left\{ \sum_{j=1}^{p+q} \psi \left(\frac{v_{l_k}^{(t)} + 1 - j}{2} \right) + (p+q) \ln 2 + \ln |A_{l_k}^{(t)}| \right\} - \frac{1}{2} \left\{ (p+q) \beta_{l_k}^{(t)-1} + v_{l_k}^{(t)} (\mathbf{u}_i - \mathbf{m}_{l_k}^{(t)})^\top A_{l_k}^{(t)} (\mathbf{u}_i - \mathbf{m}_{l_k}^{(t)}) \right\} \right. \\ \left. - \frac{1}{2\sigma_{l_k}^2} \left\{ y_i^2 - 2y_i \boldsymbol{\mu}_{wl_k}^{(t)\top} \tilde{\mathbf{v}}_i + \tilde{\mathbf{v}}_i^\top (\boldsymbol{\Lambda}_{wl_k}^{(t)-1} + \boldsymbol{\mu}_{wl_k}^{(t)} \boldsymbol{\mu}_{wl_k}^{(t)\top}) \tilde{\mathbf{v}}_i \right\} + \psi(\alpha_{l_k}^{(t)}) - \psi \left(\sum_{l'_k=1}^k \alpha_{l'_k}^{(t)} \right) - \frac{p+q}{2} \ln(2\pi) - \frac{1}{2} \ln(2\pi\sigma_{l_k}^2) \right]. \quad (31) \end{aligned}$$

The function $\psi(\cdot)$ expresses the digamma function.

2. Updating $q(\mathbf{z}_{n+1,k})$: From (23), the updating formula of $q(\mathbf{z}_{n+1,k})$ is given by

$$q^{(t+1)}(\mathbf{z}_{n+1,k}) = \prod_{l_k=1}^k \varphi_{l_k}^{(t) z_{n+1,l_k}}. \quad (32)$$

Here, $\varphi_{l_k}^{(t)}$ is given by

$$\varphi_{l_k}^{(t)} = \frac{\eta_{l_k}^{(t)}}{\sum_{l'_k=1}^k \eta_{l'_k}^{(t)}}, \quad (33)$$

where

$$\begin{aligned} \eta_{l_k}^{(t)} = \exp \left[\frac{1}{2} \left\{ \sum_{j=1}^{p+q} \psi \left(\frac{v_{l_k}^{(t)} + 1 - j}{2} \right) + (p+q) \ln 2 + \ln |A_{l_k}^{(t)}| \right\} + \psi(\alpha_{l_k}^{(t)}) - \psi \left(\sum_{l'_k=1}^k \alpha_{l'_k}^{(t)} \right) \right. \\ \left. - \frac{1}{2} \left\{ (p+q) \beta_{l_k}^{(t)-1} + v_{l_k}^{(t)} (\mathbf{u}_{n+1} - \mathbf{m}_{l_k}^{(t)})^\top A_{l_k}^{(t)} (\mathbf{u}_{n+1} - \mathbf{m}_{l_k}^{(t)}) \right\} - \frac{p+q}{2} \ln(2\pi) \right]. \quad (34) \end{aligned}$$

3. Updating $q(\boldsymbol{\pi}_k)$: We define next term with respect to each $l_k = 1, \dots, k$:

$$N_{l_k}^{(t)} = \sum_{i=1}^n r_{il_k}^{(t)} + \varphi_{l_k}^{(t)}. \quad (35)$$

From (24), the updating formula of $q(\boldsymbol{\pi}_k)$ is given by

$$q(\boldsymbol{\pi}_k)^{(t+1)} = \text{Dir}(\boldsymbol{\pi}_k \mid \boldsymbol{\alpha}_k^{(t+1)}), \quad (36)$$

where $\boldsymbol{\alpha}_k^{(t+1)} := (\alpha_1^{(t+1)}, \dots, \alpha_k^{(t+1)})^\top$, $\alpha_{l_k}^{(t+1)} := N_{l_k}^{(t+1)} + \alpha_{0_k}$ for $l_k = 1, \dots, k$.

4. Updating $q(\mathbf{M}_k, \mathbf{L}_k)$: We define next terms with respect to each $l_k = 1, \dots, k$:

$$\bar{u}_{l_k}^{(t)} = \frac{1}{N_{l_k}^{(t)}} \left\{ \sum_{i=1}^n r_{il_k}^{(t)} \mathbf{u}_i + \varphi_{l_k}^{(t)} \mathbf{u}_{n+1} \right\}, \quad (37)$$

$$\mathbf{S}_{l_k}^{(t)} = \frac{1}{N_{l_k}^{(t)}} \left\{ \sum_{i=1}^n r_{il_k}^{(t)} (\mathbf{u}_i - \bar{\mathbf{u}}_{l_k}^{(t)}) (\mathbf{u}_i - \bar{\mathbf{u}}_{l_k}^{(t)})^\top + \varphi_{l_k}^{(t)} (\mathbf{u}_{n+1} - \bar{\mathbf{u}}_{l_k}^{(t)}) (\mathbf{u}_{n+1} - \bar{\mathbf{u}}_{l_k}^{(t)})^\top \right\}. \quad (38)$$

From (25), the updating formula of $q(\mathbf{M}_k, \mathbf{L}_k)$ is given by

$$q^{(t+1)}(\mathbf{M}_k, \mathbf{L}_k) = \prod_{l_k=1}^k \mathcal{N}(\boldsymbol{\mu}_{l_k} \mid \mathbf{m}_{l_k}^{(t+1)}, (\beta_{l_k}^{(t+1)} \boldsymbol{\Lambda}_{l_k})^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_{l_k} \mid \mathbf{A}_{l_k}^{(t+1)}, \nu_{l_k}^{(t+1)}). \quad (39)$$

Here, $\beta_{l_k}^{(t+1)}$, $\nu_{l_k}^{(t+1)}$, $\mathbf{m}_{l_k}^{(t+1)}$, and $\mathbf{A}_{l_k}^{(t+1)}$ are defined as

$$\beta_{l_k}^{(t+1)} = \beta_{0_k} + N_{l_k}^{(t+1)}, \quad \nu_{l_k}^{(t+1)} = \nu_{0_k} + N_{l_k}^{(t+1)},$$

$$\mathbf{m}_{l_k}^{(t+1)} = \frac{1}{\beta_{l_k}^{(t+1)}} (\beta_{l_k}^{(0)} \mathbf{m}_{0_k} + N_{l_k}^{(t+1)} \bar{\mathbf{u}}_{l_k}^{(t+1)}),$$

$$(\mathbf{A}_{l_k}^{(t+1)})^{-1} = (\mathbf{A}_{0_k})^{-1} + N_{l_k}^{(t+1)} \mathbf{S}_{l_k}^{(t+1)} + \frac{\beta_{0_k} N_{l_k}^{(t+1)}}{\beta_{0_k} + N_{l_k}^{(t+1)}} (\bar{\mathbf{u}}_{l_k}^{(t+1)} - \mathbf{m}_{0_k}) (\bar{\mathbf{u}}_{l_k}^{(t+1)} - \mathbf{m}_{0_k})^\top.$$

5. Updating $q(\mathbf{W}_k)$: From (26), the updating formula of $q(\mathbf{W}_k)$ is given by

$$q^{(t+1)}(\mathbf{W}_k) = \prod_{l_k=1}^k \mathcal{N}(\mathbf{w}_{l_k}^{(t+1)} \mid \boldsymbol{\mu}_{wl_k}^{(t+1)}, (\boldsymbol{\Lambda}_{wl_k}^{(t+1)})^{-1}). \quad (40)$$

Here, $\boldsymbol{\Lambda}_{wl_k}^{(t+1)}$ and $\boldsymbol{\mu}_{wl_k}^{(t+1)}$ are given as

$$\boldsymbol{\Lambda}_{wl_k}^{(t+1)} = \frac{1}{\sigma_{l_k}^2} \sum_{i=1}^n r_{il_k}^{(t+1)} \tilde{\mathbf{y}}_i \tilde{\mathbf{y}}_i^\top + \boldsymbol{\Lambda}_{w0_k},$$

$$\boldsymbol{\mu}_{wl_k}^{(t+1)} = (\boldsymbol{\Lambda}_{wl_k}^{(t+1)})^{-1} \left(\frac{1}{\sigma_{l_k}^2} \sum_{i=1}^n r_{il_k}^{(t+1)} y_i \tilde{\mathbf{y}}_i + \boldsymbol{\Lambda}_{w0_k} \boldsymbol{\mu}_{w0_k} \right).$$

6. Testing convergence: We use *variational lower bound* to test the convergence and decide when to stop the iteration, since the variational Bayes algorithm guarantees the variational lower bound to increase in each iteration [16].

Using the variational distribution of each parameter or latent variable after t iterations, state $q^{(t)}(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k)$ as

$$q^{(t)}(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k) := q^{(t)}(\mathbf{z}_k^n) q^{(t)}(\mathbf{z}_{n+1:k}) q^{(t)}(\boldsymbol{\pi}_k) q^{(t)}(\mathbf{M}_k, \mathbf{L}_k) q^{(t)}(\mathbf{W}_k).$$

Then, the variational lower bound at iteration t is defined as

$$\mathcal{L}_k^{(t)} := \sum_{\mathbf{z}_k^{n+1}} \int_{\Theta_k} q^{(t)}(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k) \ln \frac{p(\mathbf{D}, \mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k)}{q^{(t)}(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k)} d\boldsymbol{\theta}_k + \ln k! \quad (41)$$

$$= E[\ln p(\mathbf{D}, \mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k)] - E[\ln q^{(t)}(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k)] \quad (42)$$

$$\begin{aligned} &= E[\ln p(\mathbf{u}^{n+1} | \mathbf{z}_k^{n+1}, \mathbf{M}_k, \mathbf{L}_k)] + E[\ln p(\mathbf{y}^n | \mathbf{v}^{n+1}, \mathbf{W}_k, \mathbf{z}_k^{n+1})] \\ &\quad + E[\ln p(\mathbf{z}_k^{n+1} | \boldsymbol{\pi}_k)] + E[\ln p(\mathbf{M}_k, \mathbf{L}_k)] \\ &\quad + E[\ln p(\boldsymbol{\pi}_k)] + E[\ln p(\mathbf{W}_k)] - E[\ln q^{(t)}(\mathbf{z}_k^n)] - E[\ln q^{(t)}(\mathbf{z}_{n+1:k})] \\ &\quad - E[\ln q^{(t)}(\mathbf{M}_k, \mathbf{L}_k)] - E[\ln q^{(t)}(\boldsymbol{\pi}_k)] - E[\ln q^{(t)}(\mathbf{W}_k)]. \end{aligned} \quad (43)$$

Note here that all the expectation in (43) is taken by distribution $q^{(t)}(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k)$. Each term in (43) can be calculated as in Appendix B. By calculating the variational lower bound at each iteration, we judge the convergence of the algorithm.

4.3 Adopting Approximated Posterior Distribution

Let T be the number of iteration when the variational lower bound satisfied the convergence condition. Then, we can derive the variational distribution of the cluster number k ($k = 1, \dots, k_{\max}$) as below:

$$q^{(T)}(k) := \frac{p(k) \exp \left\{ \mathcal{L}_k^{(T)} \right\}}{\sum_{k'=1}^{k_{\max}} p(k') \exp \left\{ \mathcal{L}_{k'}^{(T)} \right\}}. \quad (44)$$

By substituting approximated posterior distribution $q^{(T)}(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k, k) := q^{(T)}(k) q^{(T)}(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k)$ to the posterior distribution $p(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k, k | \mathbf{D})$, we can obtain (14) in the approximated form (the detail is in Appendix C) :

$$p^*(\mathbf{y}_{n+1} | \mathbf{D}) \approx \sum_{k=1}^{k_{\max}} q^{(T)}(k) \sum_{l_k=1}^k \varphi_{l_k}^{(T)} \times \mathcal{N} \left(\mathbf{y}_{n+1} \left| \boldsymbol{\mu}_{wl_k}^{(T)\top} \tilde{\mathbf{v}}_{n+1}, \frac{1}{\sigma_{l_k}^2} + \tilde{\mathbf{v}}_{n+1}^\top \left(\boldsymbol{\Lambda}_{wl_k}^{(T)} \right)^{-1} \tilde{\mathbf{v}}_{n+1} \right. \right). \quad (45)$$

Using this approximated predictive distribution, we can finally derive the next proposition.

Proposition 4.3 *Adopting the variational distribution derived by variational Bayes algorithm, the optimal prediction under Bayes criterion (13) can be approximated as*

$$\delta^*(\mathbf{D}) \approx \sum_{k=1}^{k_{\max}} q^{(T)}(k) \left(\sum_{l_k=1}^k \varphi_{l_k}^{(T)} \boldsymbol{\mu}_{wl_k}^{(T)\top} \tilde{\mathbf{v}}_{n+1} \right). \quad (46)$$

The proof of Proposition 4.3 is in Appendix D.

Note here that the term in brackets of (46) represents the approximation of the optimal prediction under Bayes criterion when fixing k as the cluster number. The prediction (46) is the weighted sum of the prediction at each cluster number, where the weight is $q^{(T)}(k)$ for each k ($k = 1, \dots, k_{\max}$).

5 Experiments

To demonstrate the behavior of our algorithm, we used various types of data. First, we generated synthetic data from the model we assumed in Sect. 2, and adopted our algorithms. In contrast, we did not only generated parameters using the priors we defined in Sect. 4.1, but also fixed the parameters to generate the data with specific tendency. In addition, we used real data to confirm that our algorithm works well in practical use, too.

For all the experiment, we set $q = 0$ which means that there is no intersection between clustering explanatory variable and regression explanatory variable. Hence, it would be easier to understand the behavior of each role of explanatory variables.

All the codes used in the experiments are opened in Github [17].

5.1 Experiment 1

5.1.1 Setting

In Experiment 1, we generated both the parameters and the data from the model described in previous sections, and checked the average of squared-error³ occurred by the prediction (46).

The procedure of Experiment 1 is shown in the following: First, we generated the cluster number k from the prior (15), and generated the hyper-parameters as follows. α_{0_k} : generated from uniform distribution on $[0, 1)$, β_{0_k} : generated from uniform distribution on $[0.001, 1.001)$, $A_{0_k} = \mathbf{I}$ (Identity matrix), ν_{0_k} : generated from uniform distribution on $[2, 3)$, $\boldsymbol{\mu}_{w0_k} = \mathbf{0}$, $\boldsymbol{\Lambda}_{w0_k} = \mathbf{I}$, and $\sigma_{l_k}^2 = 0.5$ for all $l_k = 1, \dots, k$. Next, we

³ Hereafter, we use the words “the prediction error” in short instead of using “the average of squared-error”.

Fig. 2 The prediction error of proposed algorithm and random forest

Fig. 3 The prediction error at each cluster number

generated the parameters from the priors (16)–(18). Consequently, we generated N sets of learning data and M sets of test data from the model. Changing the number of training data n from 10, 20, $\dots$, N , we calculated the prediction and the prediction error. We generated the initial values of hyper-parameters at each cluster number k in the same way as we did for genuine hyper-parameters. We repeated data generation for P times per parameter generation, parameter generation for Q times per cluster number generation, and cluster number generation for R times.

Here, each constant was stated as following: $p = 3$, $q = 0$, $r = 5$, $k_{\max} = 5$, $N = 100$, $M = 100$, $P = 100$, $Q = 10$, and $R = 10$. The iteration was stopped either when $\mathcal{L}_k^{(t+1)} - \mathcal{L}_k^{(t)} < 0.0001$ or when the iteration number reached to 30 times.

5.1.2 Result and Discussion

The result of Experiment 1 is shown in Figs. 2 and 3.

Figure 2 shows the prediction error occurred by both our algorithm and random forest algorithm [18]. Since our algorithm is the approximation of Bayes optimal prediction, it is expected to have smaller prediction error compared to any other prediction algorithms. As shown in Fig. 2, the proposed algorithm obtained expected result, where the prediction error is lower than that of random forest regardless of the number of data n . Therefore, in spite of the approximation by variational Bayes algorithm, the prediction by proposed algorithm still indicates the characteristic of the optimal prediction under Bayes criterion.

In Fig. 3, lines $k = 1, 2, 3, 4, 5$ indicates the prediction error at each cluster number, which corresponds to the term in brackets of (46). These lines can be regarded as the prediction error when the cluster number is fixed at k . The graph shows that the prediction error of our algorithm is smaller than any other prediction errors. Therefore, we should not assume the concrete cluster number so that we can obtain a better prediction.

5.2 Experiment 2

5.2.1 Setting

In Experiment 2, the cluster number and the parameters are fixed, so that we can confirm the behaviors of our algorithm when treating the data with specific tendency. We set $k = 3$, $p = 3$, $q = 0$, $r = 5$, and compared the following patterns.

- (a) Neither any clusters nor any regression lines are overlapping.

$$\begin{aligned}\mu_1 &= (2, 0, 0, 0, 0, 0, 0, 0)^T, \\ \mu_2 &= (0, 2, 0, 0, 0, 0, 0, 0)^T, \\ \mu_3 &= (0, 0, 2, 0, 0, 0, 0, 0)^T, \\ w_1 &= (0, 0, 0, 1, 2, 0, 0, 0)^T, \\ w_2 &= (0, 0, 0, 0, 0, 1, 2, 0)^T, \\ w_3 &= (0, 0, 0, 0, 0, 0, 1, 2)^T.\end{aligned}$$

- (b) Cluster 1 and 2 are overlapping.

$$\begin{aligned}\mu_1 &= (4/5, 0, 0, 0, 0, 0, 0, 0)^T, \\ \mu_2 &= (0, 4/5, 0, 0, 0, 0, 0, 0)^T, \\ \mu_3 &= (0, 0, 2, 0, 0, 0, 0, 0)^T, \\ w_1 &= (0, 0, 0, 1, 2, 0, 0, 0)^T, \\ w_2 &= (0, 0, 0, 0, 0, 1, 2, 0)^T, \\ w_3 &= (0, 0, 0, 0, 0, 0, 1, 2)^T.\end{aligned}$$

Fig. 4 The prediction error in each condition (a–d)

(c) Regression line 1 and 2 are overlapping.

$$\mu_1 = (2, 0, 0, 0, 0, 0, 0, 0)^T,$$

$$\mu_2 = (0, 2, 0, 0, 0, 0, 0, 0)^T,$$

$$\mu_3 = (0, 0, 2, 0, 0, 0, 0, 0)^T,$$

$$w_1 = (0, 0, 0, 1, 2, 0, 0, 0)^T,$$

$$w_2 = (0, 0, 0, 1, 2, 0, 0, 0)^T,$$

$$w_3 = (0, 0, 0, 0, 0, 0, 0, 1, 2)^T.$$

(d) Cluster 1 and 2 are overlapping, and regression line 1 and 2 are overlapping.

$$\mu_1 = (4/5, 0, 0, 0, 0, 0, 0, 0)^T,$$

$$\mu_2 = (0, 4/5, 0, 0, 0, 0, 0, 0)^T,$$

$$\mu_3 = (0, 0, 2, 0, 0, 0, 0, 0)^T,$$

$$w_1 = (0, 0, 0, 1, 2, 0, 0, 0)^T,$$

$$w_2 = (0, 0, 0, 1, 2, 0, 0, 0)^T,$$

$$w_3 = (0, 0, 0, 0, 0, 0, 0, 1, 2)^T.$$

For other parameters, we fixed $\Lambda_{l_k} = 10I$, $\sigma_{l_k}^2 = 0.5$ ($l_k = 1, 2, 3$). Constants and initial values for variational Bayes algorithm were set as we did in Experiment 1.

In Experiment 2, we examined the behavior of the following values:

- (i) The prediction error occurred by the prediction (46).
- (ii) The approximated posterior distribution of the cluster number K .

In the procedure of Experiment 2, we conducted the same process done in Experiment 1, except that we fixed the cluster number k in the case of Experiment 2.

5.2.2 Result and Discussion

The results of Experiment 2 are shown in Figs. 4 and 5.

Fig. 5 The approximated posterior distribution of the cluster number K . The x-axis represents the number of samples n , and the y-axis represents the approximated posterior probability of $k = 1, 2, \dots, 5$

Since condition (a) can be regarded as the benchmark, we give discussion for conditions (b–d).

For condition (b), the prediction error is larger than that of condition (a) as shown in Fig. 4. In contrast, as shown in Fig. 5, the approximated posterior distribution of the cluster number K is similar to that of condition (a). Therefore, the cause of larger prediction error can be regarded as the complexity of clustering which yields the prediction error at each cluster number k .

For condition (c), as shown in Fig. 4, the prediction error is smaller than that in condition (a) when the number of data n is small. In addition, from Fig. 5, there exists a certain value of the approximated posterior distribution of the cluster number $k = 2$ when n is small. Hence, when there is a small number of data, our algorithm indicates that it is appropriate to weight the prediction regarding the cluster number as $k = 2$ to make the optimal prediction under Bayes criterion. Conversely, when there is enough number of data n to distinguish the cluster 1 and 2, our algorithm indicates that it is suitable to weight the prediction regarding the actual cluster number $k = 3$. Therefore, we could confirm the feature of our algorithm that it automatically optimizes the cluster number k .

For condition (d), the behaviors indicated in condition (c) become more obvious. Figure 5 shows that the approximated posterior distribution of $k = 2$ is the highest at any number of data n . Since there are essentially two clusters in condition (d), the algorithm adopts $k = 2$ even when there is a large number of data.

To conclude, from Experiment 2, we confirmed the following features of our algorithm. Firstly, the complexity of clustering is one of the causes which makes the prediction error larger. Secondly, as long as we provide the appropriate $k_{\max}$, the proposed algorithm automatically tunes the weight of cluster number k for the optimal prediction under the Bayes criterion.

Fig. 6 The prediction error of our algorithm. The blue and orange line shows the error when using average blood pressure and BMI respectively as cluster explanatory variable. The error bar represents the standard error of means

5.3 Experiment 3

5.3.1 Setting

In Experiment 3, we used real data taken from Scikit-learn datasets library [19] which include explanatory variables such as age, body mass index (BMI), average blood pressure, and six blood serum measurements obtained for each of patients, as well as the target variable, a quantitative measure of disease progression one year after baseline.

In this experiment, we set the explanatory variable in several patterns, and compare the prediction error made on each generative model. Setting the quantitative measure of disease progression as target variable $y \in \mathbb{R}$, we fixed the six blood serum measurements as regression explanatory variable. Using blood pressure for cluster explanatory variable is one case, and we used BMI for cluster explanatory variable for another case.

The procedure of Experiment 3 is shown as follows: First, we randomly sampled $(N + M)$ data from the whole dataset, and divided them to N training data, and M test data. Changing the number of training data n from 10, 20, ..., N , we calculated the prediction and the prediction error. We repeated the random sampling of the data for P times. Here, we set $N = 50, M = 50, P = 100$. Constants and initial values for variational Bayes algorithm were set as we did in Experiment 1.

5.3.2 Result and Discussion

The result of Experiment 3 is shown in Fig. 6.

As can be seen from Fig. 6, using BMI as cluster explanatory variable lessen the prediction error compared to using blood pressure for cluster explanatory variable. Therefore, we can infer that the data structure which use BMI as cluster explanatory

variable gives better explanation to the given data. Hence, using our algorithm enables data analysts to understand well about the structure of the data.

6 Conclusion

In this paper, we considered the data generation model composed of several linear regression models and proposed the optimal prediction under Bayes criterion. In addition, we created the algorithm which approximately calculates the optimal prediction. Some experiments indicated the following behaviors of our proposed algorithm:

1. By taking the weighted average of prediction by the approximated posterior distribution of cluster number K , the average of prediction error becomes smaller.
2. The complexity of clustering makes the prediction error larger.
3. The algorithm automatically optimizes the weight of cluster number k , so that it realizes the approximative Bayes optimal prediction.
4. Adopting our algorithm enables us to estimate the structure of data generation.

To extend our research in the future, we will give focus on automatic selection of cluster explanatory variables and regression explanatory variables. This extension will enable use to adopt the algorithm to any kind of real value data without any tunings beforehand.

Appendix A

Proof of Proposition 3.1

By applying Bayes' theorem, the Bayes risk of decision rule $\delta(\mathbf{D})$ can be transformed as

$$\begin{aligned} \text{BR}(\delta(\cdot)) = \int_{\mathcal{D}} \left\{ \int_{\mathcal{Y}} (y_{n+1} - \delta(\mathbf{D}))^2 \sum_{k=1}^{k_{\max}} \sum_{\mathcal{Z}_k^{n+1}} \right. \\ \left. \int_{\Theta_k} p(y_{n+1} \mid \mathbf{x}_{n+1}, \mathbf{z}_k^{n+1}, \theta_k, k) p(\mathbf{z}_k^{n+1}, \theta_k, k \mid \mathbf{D}) d\theta_k dy_{n+1} \right\} \times p(\mathbf{D}) d\mathbf{D}. \end{aligned} \quad (47)$$

Therefore, minimizing the Bayes risk is equivalent to minimizing the terms in the brackets of (47). Let a represent $\delta(\mathbf{D})$, and $F(a)$ represent the terms in the brackets of (47), i.e.,

$$F(a) \equiv \int_{\mathcal{Y}} (y_{n+1} - a)^2 \sum_{k=1}^{k_{\max}} \sum_{\mathcal{Z}_k^{n+1}} \int_{\Theta_k} p(y_{n+1} | \mathbf{x}_{n+1}, \mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k, k) p(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k, k | \mathbf{D}) d\boldsymbol{\theta}_k dy_{n+1}.$$

By taking partial derivative of $F(a)$ with respect to a and setting it to 0, we obtain the minimum value of $F(a)$ when a is represented as

$$a = \int_{\mathcal{Y}} y_{n+1} p^*(y_{n+1} | \mathbf{D}) dy_{n+1}, \quad (48)$$

where

$$p^*(y_{n+1} | \mathbf{D}) := \sum_{k=1}^{k_{\max}} \sum_{\mathcal{Z}_k^{n+1}} \int_{\Theta_k} p(y_{n+1} | \mathbf{D}, \mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k, k) p(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k, k | \mathbf{D}) d\boldsymbol{\theta}_k. \quad (49)$$

Appendix B

Calculation of Variational Lower Bound

Each term in variational lower bound (43) can be written as following.

$$E[\ln p(\mathbf{u}^{n+1} | \mathbf{z}_k^{n+1}, \mathbf{M}_k, \mathbf{L}_k)] \quad (50)$$

$$\begin{aligned} &= \frac{1}{2} \sum_{l_k=1}^k N_{l_k}^{(t)} \left\{ \sum_{j=1}^{p+q} \psi \left(\frac{v_{l_k}^{(t)} + 1 - j}{2} \right) + (p+q) \ln 2 + \ln |\mathbf{A}_{l_k}^{(t)}| - (p+q) \beta_{l_k}^{(t)-1} - v_{l_k}^{(t)} \text{tr}(\mathbf{S}_{l_k}^{(t)} \mathbf{A}_{l_k}^{(t)}) \right. \\ &\quad \left. - v_{l_k}^{(t)} (\bar{\mathbf{u}}_{l_k}^{(t)} - \mathbf{m}_{l_k}^{(t)})^{\top} \mathbf{A}_{l_k}^{(t)} (\bar{\mathbf{u}}_{l_k}^{(t)} - \mathbf{m}_{l_k}^{(t)}) - (p+q) \ln(2\pi) \right\}, \\ &E[\ln p(y^n | \mathbf{v}_k^{n+1}, \mathbf{W}_k, \mathbf{z}_k^{n+1})] \end{aligned} \quad (51)$$

$$\begin{aligned} &= -\frac{1}{2} \sum_{l_k=1}^k \left[N_{l_k}^{(t)} \ln(2\pi\sigma_{l_k}^2) + \frac{1}{\sigma_{l_k}^2} \sum_{i=1}^n r_{il_k}^{(t)} \left\{ y_i^2 - 2y_i \boldsymbol{\mu}_{wl_k}^{(t)\top} \bar{\mathbf{v}}_i + \bar{\mathbf{v}}_i^{\top} (\boldsymbol{\Lambda}_{wl_k}^{-1} + \boldsymbol{\mu}_{wl_k}^{(t)} \boldsymbol{\mu}_{wl_k}^{(t)\top}) \bar{\mathbf{v}}_i \right\} \right], \\ &E[\ln p(\mathbf{z}_k^{n+1} | \boldsymbol{\pi}_k)] = \sum_{l_k=1}^k N_{l_k}^{(t)} \left\{ \psi(\alpha_{l_k}^{(t)}) - \psi \left(\sum_{l_k=1}^k \alpha_{l_k}^{(t)} \right) \right\}, \end{aligned} \quad (52)$$

$$E[\ln p(\mathbf{M}_k, \mathbf{L}_k)] = \frac{1}{2} \sum_{l_k=1}^k \left[(p+q) \ln(\beta_{0_k}/2\pi) + \sum_{j=1}^{p+q} \psi \left(\frac{v_{l_k}^{(t)} + 1 - j}{2} \right) + (p+q) \ln 2 + \ln |\mathbf{A}_{l_k}^{(t)}| \right] \quad (53)$$

$$\begin{aligned}
& -\frac{(p+q)\beta_{0_k}}{\beta_{l_k}^{(t)}} - \beta_{0_k} v_{l_k}^{(t)} (\mathbf{m}_{l_k}^{(t)} - \mathbf{m}_{0_k})^\top \mathbf{A}_{l_k}^{(t)} (\mathbf{m}_{l_k}^{(t)} - \mathbf{m}_{0_k}) - v_{0_k} \ln |\mathbf{A}_{0_k}| - v_{0_k} (p+q) \ln 2 \\
& - \frac{(p+q)(p+q-1)}{2} \ln \pi - 2 \sum_{j=1}^{p+q} \ln \Gamma \left(\frac{v_{0_k} + 1 - j}{2} \right) \\
& + (v_{0_k} - p - q - 1) \left\{ \sum_{j=1}^{p+q} \psi \left(\frac{v_{l_k}^{(t)} + 1 - j}{2} \right) + (p+q) \ln 2 + \ln |\mathbf{A}_{l_k}^{(t)}| \right\} - v_{l_k}^{(t)} \text{tr}(\mathbf{A}_{0_k}^{-1} \mathbf{A}_{l_k}^{(t)}) \Bigg], \\
E[\ln p(\boldsymbol{\pi}_k)] &= \ln \frac{\Gamma(\sum_{l_k=1}^k \alpha_{0_k})}{\prod_{l_k=1}^k \Gamma(\alpha_{0_k})} + (\alpha_{0_k} - 1) \sum_{l_k=1}^k \left\{ \psi(\alpha_{l_k}^{(t)}) - \psi \left(\sum_{l_k=1}^k \alpha_{l_k}^{(t)} \right) \right\}, \tag{54}
\end{aligned}$$

$$\begin{aligned}
E[\ln p(\mathbf{W}_k)] &= \frac{1}{2} \sum_{l_k=1}^k \left\{ -(q+r+1) \log(2\pi) + \ln |\boldsymbol{\Lambda}_{w0_k}| - \text{tr}(\boldsymbol{\Lambda}_{w0_k} (\boldsymbol{\Lambda}_{wl_k}^{(t)})^{-1} + \boldsymbol{\mu}_{wl_k}^{(t)} \boldsymbol{\mu}_{wl_k}^{(t)\top}) \right. \\
& \left. + 2 \boldsymbol{\mu}_{wl_k}^{(t)\top} \boldsymbol{\Lambda}_{w0_k} \boldsymbol{\mu}_{w0_k} - \boldsymbol{\mu}_{w0_k}^\top \boldsymbol{\Lambda}_{w0_k} \boldsymbol{\mu}_{w0_k} \right\}, \tag{55}
\end{aligned}$$

$$E[\ln q^{(t)}(\mathbf{z}_k^n)] = \sum_{i=1}^n \sum_{l_k=1}^k r_{il_k}^{(t)} \ln r_{il_k}^{(t)}, \tag{56}$$

$$E[\ln q^{(t)}(\mathbf{z}_{n+1:k})] = \sum_{l_k=1}^k \varphi_{l_k}^{(t)} \ln \varphi_{l_k}^{(t)}, \tag{57}$$

$$E[\ln q^{(t)}(\mathbf{M}_k, \mathbf{L}_k)] = \sum_{l_k=1}^k \left[\frac{1}{2} \left\{ \sum_{j=1}^{p+q} \psi \left(\frac{v_{l_k}^{(t)} + 1 - j}{2} \right) + (p+q) \ln 2 + \ln |\mathbf{A}_{l_k}^{(t)}| \right\} \right. \tag{58}$$

$$\begin{aligned}
& \left. + \frac{p+q}{2} \ln \left(\frac{\beta_{l_k}^{(t)}}{2\pi} \right) - \frac{p+q}{2} - H[q^{(t)}(\boldsymbol{\Lambda}_{l_k}^{(t)})] \right], \\
E[\ln q^{(t)}(\boldsymbol{\pi}_k)] &= \ln \frac{\Gamma(\sum_{l_k=1}^k \alpha_{l_k}^{(t)})}{\prod_{l_k=1}^k \Gamma(\alpha_{l_k}^{(t)})} \\
& + \sum_{l_k=1}^k (\alpha_{l_k}^{(t)} - 1) \left\{ \psi(\alpha_{l_k}^{(t)}) - \psi \left(\sum_{l_k=1}^k \alpha_{l_k}^{(t)} \right) \right\}, \tag{59}
\end{aligned}$$

$$E[\ln q^{(i)}(\mathbf{W}_k)] = \frac{1}{2} \sum_{l_k=1}^k \left[-(q+r+1) \ln(2\pi) + \ln |\Lambda_{\mathbf{w}_{l_k}}| - \text{tr} \left(\Lambda_{\mathbf{w}_{l_k}}^{(i)-1} + \boldsymbol{\mu}_{\mathbf{w}_{l_k}}^{(i)} \boldsymbol{\mu}_{\mathbf{w}_{l_k}}^{(i)\top} \right) + \boldsymbol{\mu}_{\mathbf{w}_{l_k}}^{(i)\top} \Lambda_{\mathbf{w}_{l_k}}^{(i)} \boldsymbol{\mu}_{\mathbf{w}_{l_k}}^{(i)} \right]. \quad (60)$$

Here, $\psi(\cdot)$ represents digamma function, and $H[\cdot]$ represents entropy of the Wishart distribution.

Appendix C

Proof of Equation (45)

By substitution of approximated posterior distribution $q^{(T)}(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k, k)$ to the posterior distribution $p(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k, k | \mathbf{D})$ in (14), we can calculate as

$$\begin{aligned} p^*(y_{n+1} | \mathbf{D}) &= \sum_{k=1}^{k_{\max}} \sum_{\mathbf{z}_k^{n+1}} \int_{\Theta_k} p(y_{n+1} | \mathbf{D}, \mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k, k) p(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k, k | \mathbf{D}) d\boldsymbol{\theta}_k \end{aligned} \quad (61)$$

$$\approx \sum_{k=1}^{k_{\max}} \sum_{\mathbf{z}_k^{n+1}} \int_{\Theta_k} p(y_{n+1} | \mathbf{D}, \mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k, k) q^{(T)}(\mathbf{z}_k^{n+1}, \boldsymbol{\theta}_k, k) d\boldsymbol{\theta}_k \quad (62)$$

$$= \sum_{k=1}^{k_{\max}} \sum_{\mathbf{z}_k} \int_{\mathcal{W}_k} p(y_{n+1} | \mathbf{v}_{n+1}, \mathbf{z}_{n+1, k}, \mathbf{W}_k, k) q^{(T)}(\mathbf{z}_{n+1, k}) q^{(T)}(\mathbf{W}_k) q^{(T)}(k) d\mathbf{W}_k \quad (63)$$

$$\begin{aligned} &= \sum_{k=1}^{k_{\max}} q^{(T)}(k) \sum_{\mathbf{z}_k} \int_{\mathcal{W}_k} \prod_{l_k=1}^k \mathcal{N}(y_{n+1} | \mathbf{w}_{l_k}^{(T)\top} \tilde{\mathbf{v}}_{n+1}, \sigma_{l_k}^2)^{z_{n+1, l_k}} \\ &\quad \cdot \prod_{l_k=1}^k \varphi_{l_k}^{(T) z_{n+1, l_k}} \cdot \prod_{l_k=1}^k \mathcal{N}(\mathbf{w}_{l_k} | \boldsymbol{\mu}_{\mathbf{w}_{l_k}}^{(T)}, \Lambda_{\mathbf{w}_{l_k}}^{(T)-1}) d\mathbf{W}_k \end{aligned} \quad (64)$$

$$= \sum_{k=1}^{k_{\max}} q^{(T)}(k) \sum_{\mathbf{z}_k} \int_{\mathcal{W}_k} \prod_{l_k=1}^k \left[\left\{ \varphi_{l_k}^{(T)} \mathcal{N}(y_{n+1} | \mathbf{w}_{l_k}^{(T)\top} \tilde{\mathbf{v}}_{n+1}, \sigma_{l_k}^2) \right\}^{z_{n+1, l_k}} \cdot \mathcal{N}(\mathbf{w}_{l_k} | \boldsymbol{\mu}_{\mathbf{w}_{l_k}}^{(T)}, \Lambda_{\mathbf{w}_{l_k}}^{(T)-1}) \right] d\mathbf{W}_k. \quad (65)$$

Here, define l_k^* as l_k where $z_{n+1, l_k} = 1$, so that we can rewrite (65) as

$$\sum_{k=1}^{k_{\max}} q^{(T)}(k) \sum_{l_k^* \neq 1}^k \int_{\mathcal{W}} \left\{ \varphi_{l_k^*}^{(T)} \mathcal{N}\left(y_{n+1} \mid \mathbf{w}_{l_k^*}^{(T)\top} \tilde{\mathbf{v}}_{n+1}, \sigma_{l_k^*}^2\right) \cdot \mathcal{N}(\mathbf{w}_{l_k^*} \mid \boldsymbol{\mu}_{\mathbf{w}_{l_k^*}}^{(T)}, \boldsymbol{\Lambda}_{\mathbf{w}_{l_k^*}}^{(T)-1}) \cdot \prod_{l_k \neq l_k^*} \mathcal{N}(\mathbf{w}_{l_k} \mid \boldsymbol{\mu}_{\mathbf{w}_{l_k}}^{(T)}, \boldsymbol{\Lambda}_{\mathbf{w}_{l_k}}^{(T)-1}) \right\} d\mathbf{w}_k \quad (66)$$

$$= \sum_{k=1}^{k_{\max}} q^{(T)}(k) \sum_{l_k^* \neq 1}^k \varphi_{l_k^*}^{(T)} \mathcal{N}\left(y_{n+1} \mid \boldsymbol{\mu}_{\mathbf{w}_{l_k^*}}^{(T)\top} \tilde{\mathbf{v}}_{n+1}, \frac{1}{\sigma_{l_k^*}^2} + \tilde{\mathbf{v}}_{n+1}^\top \left(\boldsymbol{\Lambda}_{\mathbf{w}_{l_k^*}}^{(T)}\right)^{-1} \tilde{\mathbf{v}}_{n+1}\right) \prod_{l_k \neq l_k^*} \underbrace{\int_{\mathcal{W}_{l_k}} \mathcal{N}(\mathbf{w}_{l_k} \mid \boldsymbol{\mu}_{\mathbf{w}_{l_k}}^{(T)}, \boldsymbol{\Lambda}_{\mathbf{w}_{l_k}}^{(T)-1}) d\mathbf{w}_{l_k}}_{=1} \quad (67)$$

$$= \sum_{k=1}^{k_{\max}} q^{(T)}(k) \sum_{l_k^* \neq 1}^k \varphi_{l_k^*}^{(T)} \mathcal{N}\left(y_{n+1} \mid \boldsymbol{\mu}_{\mathbf{w}_{l_k^*}}^{(T)\top} \tilde{\mathbf{v}}_{n+1}, \frac{1}{\sigma_{l_k^*}^2} + \tilde{\mathbf{v}}_{n+1}^\top \left(\boldsymbol{\Lambda}_{\mathbf{w}_{l_k^*}}^{(T)}\right)^{-1} \tilde{\mathbf{v}}_{n+1}\right). \quad (68)$$

Appendix D

Proof of Proposition 4.3

As proven in Proposition 3.1, optimal prediction under Bayes criterion is mean of the predictive distribution $p^*(y_{n+1} \mid \mathbf{D})$. Here, we use approximated predictive distribution which we derived in (45), and take the mean of it. Then, we can obtain the prediction $\delta^*(\mathbf{D})$ in approximated form as

$$\delta^*(\mathbf{D}) \approx \sum_{k=1}^{k_{\max}} q^{(T)}(k) \left(\sum_{l_k=1}^k \varphi_{l_k}^{(T)} \boldsymbol{\mu}_{\mathbf{w}_{l_k}}^{(T)\top} \tilde{\mathbf{v}}_{n+1} \right). \quad (69)$$

Acknowledgements The authors are very grateful to the reviewers for their helpful comments.

Author Contributions Conceptualization: HM, SS, YN, YI and TM; Methodology: HM, SS, YN, YI and TM; Software: HM, YN, and YI; Validation: HM, SS, YN, YI and TM; Formal analysis: HM, SS, YN, YI and TM; Investigation: HM, SS, YN, YI and TM; Resources: HM; Data curation: HM; Writing—original draft preparation: HM, SS, YN; Writing—review & editing: HM, SS, YN, YI and TM; Visualization: HM; Supervision: TM; Project administration: TM; Funding acquisition: SS and TM. All authors have read and agreed to the published version of the manuscript.

Funding This work was supported by JSPS KAKENHI Grant Numbers JP17K06446, JP18K11585, JP19K04914, JP19K14989.

Availability of Data and Material See [17] for the code used for experiments. The data sets used in Experiment 3 is available at [19].

Declarations

Conflict of Interests The authors declare they have no competing interests.

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Nathan, G., Holt, D.: The effect of survey design on regression analysis. *J. R. Stat. Soc. Ser. B (Methodological)* **42**(3), 377–386 (1980)
2. Jacobs, R.A., Jordan, M.: Adaptive mixture of local experts. *Neural Comput.* **3**(1), 78–88 (1991)
3. Jiang, Y., Conglian, Y., Qinghua, J.: Model selection for the localized mixture of experts models. *J. Appl. Stat.* **45**(11), 1994–2006 (2018)
4. Bishop, C.M., Svensen, M.: Bayesian hierarchical mixtures of experts. In: *Uncertainty in Artificial Intelligence: Proceedings of the Nineteenth Conference*, (2003)
5. Baldacchino, T., Worden, K., Rowson, J.: Robust nonlinear system identification: Bayesian mixture of experts using the t-distribution. *Mech. Syst. Signal Process.* **85**, 977–992 (2017)
6. Mossavat, I., Amft, O.: Sparse Bayesian hierarchical mixture of experts. In: *IEEE Statistical Signal Processing Workshop (SSP)*, (2011)
7. Iikubo, Y., Horii, S., Matsushima, T.: Sparse Bayesian hierarchical mixture of experts and variational inference. In: *ISITA*, (2018)
8. Iikubo, Y., Horii, S., Matsushima, T.: Model selection of Bayesian hierarchical mixture of experts based on variational inference. In: *IEEE International Conference on Systems, Man and Cybernetics (SMC)*, (2019)
9. Nusser, S., Otte, C., Hauptmann, W.: An EM-based piecewise linear regression algorithm. In: Corchado, E., Abraham, A., Pedrycz, W. (eds.) *Hybrid Artificial Intelligence Systems*, pp. 466–474. Springer Berlin Heidelberg, Berlin (2008)
10. Eto, R., Fujimaki, R., Morinaga, S., Tamano, H.: Fully-automatic Bayesian piecewise sparse linear models. *J. Mach. Learn. Res.* **33**, 238–246 (2004)
11. Burgard, J.P., Dorr, P.: Survey-weighted generalized linear mixed models. In: *Research Papers in Economics*, No. 1/18, (2018)
12. Damas, B., Victor, J.S.: Online learning of single- and multivalued functions with an infinite mixture of linear experts. *Neural Comput.* **25**(11), 3044–3091 (2013)
13. Tan, S.L., Nott, D.J.: Variational approximation of mixtures of linear mixed model. *J. Comput. Graph. Stat.* **23**(2), 564–585 (2014)
14. Liang, J., Chen, K., Lin, M., Zhang, C., Wang, F.: Robust finite mixture regression for heterogeneous targets. *Data Min. Knowl. Disc.* **32**, 1509–1560 (2018)
15. Berger, J.O.: *Statistical Decision Theory and Bayesian Analysis*. Springer, Berlin (1985)
16. Bishop, C.M.: *Pattern Recognition and Machine Learning*. Springer, Berlin (2006)
17. Codes for experiments. [Online]. <https://github.com/syttw23/experiments.git>. Accessed 15 Sept 2021
18. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)

19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)

Authors and Affiliations

Haruka Murayama¹ · Shota Saito¹ · Yuji Iikubo¹ · Yuta Nakahara² ·
Toshiyasu Matsushima¹

Shota Saito
shota@aoni.waseda.jp

Yuji Iikubo
yuji-iikubo.8@fuji.waseda.jp

Yuta Nakahara
yuta.nakahara@aoni.waseda.jp

Toshiyasu Matsushima
toshimat@waseda.jp

¹ Department of Applied Mathematics, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

² Center for Data Science, Waseda University, Waseda-cho 27, Shinjuku-ku, Tokyo 162-0042, Japan