Hierarchical Relevance Determination based on Information Criterion Minimization

Shunsuke Hirose¹ · Tomotake Kozu¹ · Yingzi Jin¹ · Yuichi Miyamura¹

Received: 18 February 2020 / Accepted: 19 June 2020 / Published online: 3 July 2020
© The Author(s) 2020

Abstract

This paper addresses the issue of hierarchical relevance determination (HRD), which boils down to determining all degrees of freedom in a supervised mixture distribution automatically. Such relevance determination is useful for wide range of machine learning applications. However, it is difficult to solve the HRD task because its objective function includes L_0 norm such as the number of models in the mixture and the optimal features to use. Our contribution is twofold. As the main contribution, we formally defined the HRD problem and subsequently proposed a solution strategy, the SICM (sequential information criterion minimization) algorithm. The SICM algorithm enables us to continuously minimize an information criterion such as AIC or BIC, both of which includes the number of parameters, and therefore it enables us to determine all the degrees of freedom automatically. As another contribution, we realized a concrete implementation of the ideas, and tested this on actual data. Experiments using a hierarchical model which is constructed using the SICM algorithm have revealed that SICM has capability of constructing interpretable and highly accurate model.

Keywords Mixture modeling · Model selection · Variable selection · Information criterion · Continuous minimization · Interpretability

Introduction

Recently, demand has been growing for the use of interpretable prediction models in many machine learning problems, whereas uninterpretable models with high prediction accuracy such as deep learning have been widely used. For example, when constructing a credit rating model, it is legally required that the model be made interpretable. Interpretable models include decision tree and regression models such as linear regression and logistic regression.

Such frequently used interpretable models are simple. Therefore, they are usually inferior to uninterpretable models in terms of prediction accuracy. One solution for prediction accuracy improvement is to construct an interpretable model by combining multiple interpretable models.

In addition, it is important to conduct relevance determination to construct an interpretable model. Relevance determination is useful to prune irrelevant features and to lead

sparse models. Consequently, it is an important machine-learning tool.

Motivated by these backgrounds, we address the issue of relevance determination in supervised mixture distributions [1]. We designate this issue as hierarchical relevance determination (HRD). The HRD task is simultaneous optimization of the number of mixture components, model parameters of the individual components and latent variables for assigning observations to the components, and to select an optimal subset of the input variables used in the individual components.¹ As described below, the task is decomposed into a mixture model selection task and a variable selection task.

Mixture model selection task (model selection task hereinafter) is an estimation task for a mixture distribution, which is to optimize the number of mixture components, model parameters of the individual components and latent variables simultaneously. The expectation-maximization (EM) algorithm [2] has been used widely to estimate the parameters of mixture distributions. However, the number of

✉ Shunsuke Hirose
shunsuke.hirose@tohatsu.co.jp

¹ Deloitte Touche Tohmatsu LLC, Marunouchi Nijubashi Building, 3-2-3 Marunouchi, Chiyoda-ku, Tokyo, Japan

¹ There is no representative name for such a task, though it is an existing and well-known problem. Therefore we have given the name HRD in this paper.

components can not be optimized using the EM algorithm. Model comparison approaches are often used to determine the number of components automatically. A mixture with the minimum criterion value is selected as the best distribution from a set of candidate mixtures with different numbers of components. As a criterion, an information criterion such as AIC (Akaike information criterion) [3] and BIC (Bayesian information criterion) [4] has been usually used. AIC and BIC are model estimation measures, and they penalize mean training fit by the number of model parameters. Therefore, by choosing the minimum AIC/BIC model, we can determine the model complexity such as the number of components. The model comparison approaches are computationally expensive because we must estimate all candidate mixtures. We can avoid computational cost problems using Bayesian inference methods that directly minimize the upper bound of a negative marginal log-likelihood and can therefore estimate the number of components through a single mixture estimation. As the Bayesian inference methods, variational Bayesian inference (VB) [1, 5], collapsed variational Bayesian inference (CVB) [6] and factorized asymptotic Bayesian inference (FAB) [7] have been proposed. The first, VB entails the shortcoming that the upper bound is loose because the latent variables and model parameters are assumed to be independent. In the case of CVB and FAB, the independence assumption is not required. Therefore the upper bounds of CVB and FAB are tighter than that of VB.

The variable selection task is an estimation task for a single distribution, which is to optimize the parameters and to select an optimal subset of input variables. Model comparison approaches such as the forward backward algorithm [8, 9] have often been used for variable selection. As in the case of model selection, model comparison approaches are computationally expensive. Sparse estimation methods such as Lasso [10], SCAD [11] and least angle regression [12] have often been used for variable selection. The methods can solve the variable selection task because they can derive sparse parameters by minimizing the regularized objective functions. We can avoid the computational cost difficulties posed by model comparison approaches using sparse estimation methods, which directly minimize the regularized objective functions and therefore only require single estimation.

Then, how can the HRD task be solved? Apparently, we can readily construct a solution by simply combining the model selection and variable selection methods described above. However, simple combinations entail technical difficulties as discussed below. It is unrealistic to solve the HRD task using the model comparison approaches because the computational cost becomes too large. In the case of HRD, the number of candidate models becomes much larger than those in the cases of model selection and variable selection. It is also unrealistic to accomplish the task using sparse

estimation approaches because it is not trivial how to solve the model selection task using a sparse estimation method. Using VB, we can solve the HRD task. However, VB presents the following two shortcomings. First, it is difficult to apply VB widely because update equations must be derived analytically. Second, it is required that the latent variables and the model parameters must be assumed as independent, although their dependence is fundamentally important for the true distributions. We can also solve the HRD task using the FAB-based method proposed in an earlier report of the literature [13]. The method solves the model selection task through continuous optimization (by variational inference) and the variable selection task by discrete optimization (using a model comparison approach).

Using these inference methods, hierarchical models have been developed and applied to many areas. Such models include Gaussian process mixtures [14], hidden Markov mixtures [15], gamma mixtures [16], and hierarchical multinomial-Dirichlet model [17].

As described in this paper, we propose the sequential information criterion minimization (SICM) algorithm, which is a method for relevance determination in mixture distributions. In addition to being a “consistent” framework for solving the HRD task, SICM has the following properties. First, the objective function of SICM is consistent. Therefore, for example, SICM is NOT a method which solves the model selection task by VB and solves the variable selection task by Lasso. Actually, SICM solves the HRD task as a minimization problem of an information criterion. Second, it is easy to derive the model parameter update equations used in SICM. In fact, SICM estimates the model parameters by iterating L_1 -regularized sparse estimation. It is, therefore, not necessary to derive the parameter update equations analytically as in the VB case. Third, SICM require no assumption of independence between the latent variables and the model parameters. Fourth, the optimization method used in SICM is consistent. Therefore, for example, SICM is NOT a method which solves the model selection task by variational inference (by continuous optimization) and solves the variable selection task by conducting model comparison (by discrete optimization). SICM minimizes the objective function continuously using variational inference and L_1 -regularized sparse estimation.

The key ideas of SICM are summarized as follows (“Overall flow of the SICM algorithm” presents details). First, we use an information criterion, AIC or BIC, as the objective function. SICM solves the HRD task by directly minimizing an information criterion. When minimizing BIC, SICM corresponds to the Bayesian inference methods such as VB and FAB because BIC is an approximate representation of a negative marginal log-likelihood. Second, in SICM, the difficulty of the L_0 term included in an information criterion is overcome by a concave continuous approximation.

The resulting optimization problem is solved using a method based on the majorization-minimization (MM) algorithm [18]. The MM algorithm minimizes an objective function approximately by minimizing an upper bound of the objective. Third, independence between the latent variables and the model parameters is not assumed because only the marginal distribution of the latent variables is used (that of the model parameters is not used) when variational inference is used in SICM.

We propose a method for constructing an interpretable hierarchical model by the application of SICM. The model is supervised and is represented as a combination of a decision tree and regression models. When conducting prediction, we assign a regression model to an observation by selection from a set of regression models in accordance with the decision tree. By SICM, the number of regression models used in the hierarchical model and subsets of input variables used in the individual regressions are determined automatically.

We demonstrate the utility of the SICM algorithm through experiments conducted using nine UCI datasets [19]. The results indicate that we can construct an interpretable prediction model with higher prediction accuracy than frequently used interpretable models such as decision tree and logistic regression. This is true because, in binary classification problems, the above-mentioned interpretable hierarchical model based on SICM (1) outperformed decision tree and logistic regression, (2) outperformed VB, and (3) performed comparably to support vector machine (SVM), which is representative of uninterpretable models having high prediction accuracy.

In summary, the main contributions of this study are the following.

1. We propose the SICM algorithm, which solves the HRD problem by continuously minimizing an information criterion: AIC or BIC.
2. As an SICM application, we propose an interpretable hierarchical model represented as a combination of decision tree and regression models.
3. Through binary classification experiments, we demonstrate that SICM is useful for supporting construction of an interpretable model with higher prediction performance than either decision tree or regression because, in the experiments, the earlier described interpretable hierarchical model based on SICM outperformed decision tree and regression and performed comparably to SVM, an uninterpretable model that exhibits high performance.

The remainder of this paper is organized as follows. The next section provides a problem setting of HRD. The following section explains the proposed SICM algorithm, which solves the HRD problem by minimizing information

criteria continuously. The next section proposes a method for constructing an interpretable hierarchical model based on the SICM algorithm. The following section explains the experimentally obtained findings. The last section presents concluding remarks.

Problem Setting

This section presents a description of our problem settings. We consider the hierarchical relevance determination task for supervised mixture distributions, which we explain below.

Hierarchical Relevance Determination Problem

Presuming that we are given a dataset of N observations, each of which consists of D numerical input variables $\mathbf{x}$ and a target variable y :

$$(Y, X) = \{y_n, \mathbf{x}_n\}_{n=1}^N. \quad (1)$$

We regard supervised mixture distributions to be represented as

$$p(Y|X, Z, \theta) = \prod_{n=1}^N \prod_{k=1}^K (p_k(y_n|\mathbf{x}_n, \theta_k))^{z_{nk}}, \quad (2)$$

where p , p_k , K , Z , and θ_k respectively represent a probability density function (pdf) of a mixture, a pdf of the k th mixture component, the number of components, latent variables explained below, and parameters of the k th mixture. Let us denote a set of latent variables as $Z = \{z_{nk}\}$ ($n = 1, \dots, N$, $k = 1, \dots, K$), where Z is a set of binary variables representing the component assignments. If the n th observation ($y_n, \mathbf{x}_n$) is generated from the k th component distribution p_k , then $z_{nk} = 1$, and $z_{nk} = 0$ otherwise. Latent variables of one observation are mutually exclusive. Therefore, $\sum_{k=1}^K z_{nk} = 1$ holds. We assume that the latent variables follow the following distribution:

$$p(Z|\pi) = \prod_{n=1}^N \prod_{k=1}^K (\pi_k)^{z_{nk}} \left(0 \leq \pi_k \leq 1, \sum_{k=1}^K \pi_k = 1 \right). \quad (3)$$

As described herein, we consider a task of hierarchical relevance determination (HRD) for mixture distributions. The HRD task is to optimize, simultaneously, the number of mixtures K , the parameters θ , the mixing coefficients π and the latent variables Z , and to select an optimal subset of the input variables used in the individual components. The HRD task is to determine automatically all the degrees of freedom in the mixture distribution.

As discussed below, we solve the task by minimizing an information criterion. Therefore, when solving the task, we assume that the following two conditions hold. First, $\{p_k(y_n | \mathbf{x}_n, \theta_k)\}_{k=1}^K$ satisfy the regularity conditions by which the Fisher information matrices of $\{p_k(y_n | \mathbf{x}_n, \theta_k)\}_{k=1}^K$ are non-singular around the maximum likelihood estimators. Second, the optimal assignment Z is unique.

Decomposition of the Hierarchical Relevance Determination Problem

We decompose the HRD problem into two: variable selection and model selection.

Variable Selection Problem

Next we consider the HRD problem where the number of components K is fixed to 1 and the pdf is represented as

$$p(Y|X, \theta) = \prod_{n=1}^N p(y_n | \mathbf{x}_n, \theta). \quad (4)$$

We define variable selection as the relevance determination problem in this case. The variable selection task is to optimize parameter θ and to select an optimal subset of the input variables simultaneously.

Model Selection Problem

Next, we consider the HRD problem for which all the input variables are used in the individual components. We define model selection as the relevance determination problem in this case. The model selection task is to optimize, simultaneously, the number of mixtures K , the parameters θ , the mixing coefficients π , and the latent variables Z .

Proposed Method: SICM Algorithm

In this section, we propose the sequential information criterion minimization (SICM) algorithm, which solves the HRD task by continuously minimizing an information criterion of a supervised mixture distribution.

We derive the SICM algorithm for HRD by the following procedure. First, we propose an algorithm for solving the variable selection task by minimizing an information criterion continuously. Second, we propose an algorithm for solving the model selection task by continuously minimizing an information criterion. Third, we derive the SICM algorithm for solving the HRD problem by combining these two algorithms.

Overall Flow of the SICM Algorithm

We briefly review the SICM algorithm. We denote parameters included in the information criterion as Θ . In the case of variable selection, $\Theta = \theta$. In the case of model selection and HRD, $\Theta = (Z, \theta)$.

The overall flow of the SICM algorithm is summarized as presented below.

1. We use the objective function E , which satisfies $\min_{\Theta} E(\Theta) = \text{IC}^*$, where IC^* represents the minimum of an information criterion. We then formulate a minimization problem of the information criterion as a minimization problem of E .
2. We approximate the L_0 term of the parameters in the objective function by a function of L_1 norm, and numerically stabilize the problem as

$$\|\Theta\|_0 \simeq g(\|\Theta\|_1), \quad (5)$$

$$E \longrightarrow E|_{\|\Theta\|_0 \rightarrow g(\|\Theta\|_1)}. \quad (6)$$

3. We introduce new parameters Θ_0 and derive an upper bound of the objective function:

$$E(\Theta) \leq F(\Theta, \Theta_0). \quad (7)$$

4. By iteratively minimizing F , we estimate Θ as

$$\Theta^{(s+1)} = \arg \min_{\Theta} F(\Theta, \Theta_0^{(s)}), \quad (8)$$

$$\Theta_0^{(s+1)} = \Theta^{(s+1)}, \quad (9)$$

where s represents an update step. The minimization method by which an upper bound of an objective function is iteratively minimized is called the majorization-minimization (MM) algorithm [18].

Objective Function

Next, we present a definition of the objective function of SICM.

Let us denote Θ as

$$\Theta = (h, v), \quad (10)$$

where h and v respectively represent a subset of the parameters used (selected) in a model and values of the “selected” parameters. In the case of regression models, h represents a subset of input variables which have non-zero regression coefficients, and v represents the values of the non-zero coefficients. An information criterion and its minimum are expressed as

$$IC(h, v_{ML}(h)) = -\log p(Y|X, h, v_{ML}(h)) + f(h, v_{ML}(h)), \quad (11)$$

$$IC^* = \min_h IC(h, v_{ML}(h)), \quad (12)$$

where f represents a model complexity, an L_0 term included in the information criterion, and $v_{ML}(h)$ represents a maximum likelihood estimator where the subset is fixed to h :

$$v_{ML}(h) = \arg \min_v -\log p(Y|X, h, v). \quad (13)$$

Model complexity f is dependent on h and is independent of v . Therefore, f is expressed as

$$f(h, v_{ML}(h)) = f(h) = f(h, v). \quad (14)$$

From Eqs. (10) to (14), the minimum of the information criterion is expressed as

$$\begin{aligned} IC^* &= \min_h \left[\min_v -\log p(Y|X, h, v) + f(h) \right] \\ &= \min_{h,v} -\log p(Y|X, h, v) + f(h, v) \\ &= \min_{\theta} -\log p(Y|X, \theta) + f(\theta). \end{aligned} \quad (15)$$

As described in this paper, we solve the HRD problem by minimizing the information criterion of the mixture distribution. For this purpose, we use E in Eq. (17) below as the objective function to be minimized. We consider the following minimization problem:

$$\min_{\theta} E(\theta), \quad (16)$$

$$E(\theta) = -\log p(Y|X, \theta) + f(\theta). \quad (17)$$

By solving the minimization problem in Eq. (16), we can minimize the information criterion for the following reasons. The objective function E is derived by replacing v_{ML} of the information criterion with v . Therefore, E is not equal to the information criterion itself. However, the minimization problem of E coincides with that of the information criterion because the minimum E is equal to the minimum information criterion value IC^* , as shown in Eq. (15).

SICM for Variable Selection

In this section, we propose a method for solving the variable selection problem. We consider the following information criterion minimization problem:

$$\min_{\theta} E(\theta), \quad (18)$$

$$E(\theta) = -\log p(Y|X, \theta) + f(\theta), \quad (19)$$

$$f(\theta) = c_{IC} \|\theta\|_0, \quad (20)$$

$$c_{AIC} = 1, \quad (21)$$

$$c_{BIC} = \frac{1}{2} \log N. \quad (22)$$

We use AIC or BIC as the information criterion to be minimized. We set $c_{IC} = c_{AIC}$ ($c_{IC} = c_{BIC}$) when we use AIC (BIC).

It is difficult to treat $E(\theta)$ numerically because $\|\theta\|_0$ is not always continuous with respect to θ . Therefore we use the following approximation:

$$\|\theta\|_0 = \sum_i \|\theta_i\|_0 \simeq \sum_i \frac{|\theta_i|}{|\theta_i| + \eta} \quad (\eta > 0), \quad (23)$$

where θ_i and η respectively represent the i th component of θ and a user-defined small positive constant. The approximation in Eq. (23) is justified because $\|s\|_0$ and $\frac{|s|}{|s| + \eta}$ (s is a scalar) share the following properties: (1) they become 0 when $s = 0$; (2) they rapidly approach 1 when $|s|$ increases; (3) they are convex upward functions with respect to $|s|$; and (4) they become nearly equal when η approaches 0.

We introduce new parameters θ_0 , which have the same dimension as θ . $\frac{|\theta_i|}{|\theta_i| + \eta}$ in Eq. (23) is a convex upward function of $|\theta_i|$. Therefore, $\|\theta\|_0$ has an upper bound as

$$\begin{aligned} \|\theta\|_0 &\simeq \sum_i \frac{|\theta_i|}{|\theta_i| + \eta} \leq \sum_i \frac{\eta}{(|\theta_{0i}| + \eta)^2} |\theta_i| \\ &\quad + \frac{|\theta_{0i}|^2}{(|\theta_{0i}| + \eta)^2}, \end{aligned} \quad (24)$$

where θ_{0i} represents the i th component of θ_0 . The equality in Eq. (24) holds if $\theta = \theta_0$ holds because $|\theta_{0i}|$ represents a point of contact of $u(|\theta_i|) = \frac{|\theta_i|}{|\theta_i| + \eta}$ and its tangent line. Using this inequality, the following upper bound of the objective function is derived:

$$E(\theta) \leq F(\theta, \theta_0) \quad (25)$$

$$\begin{aligned} &= -\log p(Y|X, \theta) \\ &\quad + c_{IC} \sum_i \left\{ \frac{\eta}{(|\theta_{0i}| + \eta)^2} |\theta_i| + \frac{|\theta_{0i}|^2}{(|\theta_{0i}| + \eta)^2} \right\}. \end{aligned} \quad (26)$$

Using the upper bound in Eq. (26), we define the variable selection problem as the following minimization problem:

$$\min_{\theta, \theta_0} F(\theta, \theta_0). \quad (27)$$

By minimizing $F(\theta, \theta_0)$ in Eq. (26), we approximately minimize the original objective function E .

We propose an algorithm for minimizing F , i.e., for solving the variable selection task. The algorithm sequentially minimizes F . It therefore approximately minimizes the information criterion. Therefore, we designate it as the sequential information criterion minimization (SICM) algorithm.

We summarize the SICM algorithm for variable selection as Algorithm 1. The algorithm consists of the following two steps. First, F is minimized with respect to θ . As shown in Eq. (26), the problem in Eq. (27) is an L_1 -regularized maximum likelihood estimation problem with respect to θ . The

L_1 regularization term has weights $\left\{ \frac{\eta}{(|\theta_{0i}| + \eta)^2} \right\}$ corresponding to $\{|\theta_i|\}$ like that of adaptive lasso [20]. It is important that, different from adaptive lasso, the weights be determined automatically based on the information criterion. Second, F is minimized with respect to θ_0 . The solution of this problem is $\theta_0 = \theta$ because the equality in Eq. (24) holds if $\theta = \theta_0$ holds, as discussed above.

Algorithm 1 SICM algorithm for variable selection.

---

1. Setting. Choose AIC or BIC as the information criterion to be minimized. Input the user-defined parameters: η (positive constant to approximate an L_0 norm) and ϵ (threshold for convergence condition used in Step 5 below).
2. Initialization. Input the training observations, (Y, X) . Represent an update step as s and set $s = 0$. Conduct maximum likelihood estimation and define its estimator as $\theta^{(s=0)} = \theta_0^{(s=0)}$. Estimate the objective function: $F^{(s=0)} = F(\theta^{(s=0)}, \theta_0^{(s=0)})$.
3. Update of the parameters. Update θ by solving the following minimization problem.

$$\theta^{(s+1)} = \operatorname{argmin}_{\theta} F(\theta, \theta_0^{(s)}). \quad (28)$$

Update θ_0 as follows.

$$\theta_0^{(s+1)} = \operatorname{argmin}_{\theta_0} F(\theta^{(s+1)}, \theta_0) = \theta^{(s+1)}. \quad (29)$$

4. Update the objective function. Estimate the objective function as shown below.

$$F^{(s+1)} = F(\theta^{(s+1)}, \theta_0^{(s+1)}). \quad (30)$$

5. Convergence determination. If the following convergence condition

$$F^{(s)} - F^{(s+1)} < \epsilon \quad (31)$$

is satisfied, then go to Step 6. Otherwise, set $s = s + 1$ and go to Step 3.

6. Output the results and stop the algorithm.
-

SICM for Model Selection

In this section, we propose a method for solving the model selection problem. We consider the following information criterion minimization problem:

$$\min_{Z, \theta} E(Z, \theta), \quad (32)$$

$$E(Z, \theta) = -\log p(Y|X, Z, \theta) + f(Z, \theta), \quad (33)$$

$$f(Z, \theta) = \sum_{k=1}^K g_{\text{IC}}(N_k) \left(N_k = \sum_{n=1}^N z_{nk} \right), \quad (34)$$

$$g_{\text{AIC}}(N_k) = \|N_k\|_0 (N + D_k), \quad (35)$$

$$g_{\text{BIC}}(N_k) = \frac{1}{2} \|N_k\|_0 (N \log N + D_k \log N_k), \quad (36)$$

where D_k represents the degrees of freedom of the k -th mixture component $p_k(y|x, \theta_k)$ and is expressed as $D_k = \|\theta_k\|_0$. In the case of the model selection task, D_k is a constant. We set $g_{\text{IC}} = g_{\text{AIC}}$ ($g_{\text{IC}} = g_{\text{BIC}}$) when we minimize AIC (BIC).

We derive an upper bound of the objective function $E(Z, \theta)$, as in the case of variable selection. We denote the minimum of $E(Z, \theta)$ as $E(Z^*, \theta^*)$. A minimum is always less than or equal to an expectation. Consequently, the following inequality holds

$$e^{-E(Z^*, \theta^*)} \geq \langle e^{-E(Z, \theta)} \rangle_{p(Z|\pi)} = \int dZ p(Z|\pi) [p(Y|X, Z, \theta) e^{-f(Z, \theta)}] \quad (37)$$

$$= \int dZ p(Y, Z|X, \theta, \pi) e^{-f(Z, \theta)}. \quad (38)$$

Therein, $\langle \cdot \rangle_p$ represents the expectation of $\cdot$ with respect to p . By applying Jensen's inequality (see Appendix A) to $-\log$ of Eq. (38), we obtain the following upper bound as

$$E(Z^*, \theta^*) \leq -\log \int dZ q(Z) \frac{p(Y, Z|X, \theta, \pi) e^{-f(Z, \theta)}}{q(Z)} \quad (39)$$

$$\leq -\int dZ q(Z) \log \frac{p(Y, Z|X, \theta, \pi) e^{-f(Z, \theta)}}{q(Z)},$$

where $q(Z)$ represents a pdf of Z .

As in the variable selection case, we use the following approximation of the L_0 term:

$$\|N_k\|_0 = \frac{N_k}{N_k + \eta} \quad (\eta > 0), \quad (40)$$

where η is a user-defined positive constant. Also, $g(N_k)$ is a convex upward function of N_k . Consequently, by introducing a pdf $\bar{q}(Z)$, the following upper bound is derived:

$$g(N_k) \leq g(\bar{N}_k) + g'(\bar{N}_k)(N_k - \bar{N}_k), \quad (41)$$

$$g'(N_k) = \frac{\partial}{\partial N_k} g(N_k), \quad (42)$$

$$\bar{N}_k = \langle N_k \rangle_{\bar{q}(Z)} = \langle z_{nk} \rangle_{\bar{q}(Z)}. \quad (43)$$

By combining inequalities (39) and (41), the following upper bound of the objective function E is derived:

$$E(Z^*, \theta^*) \leq F(\theta, \pi, q_Z, \bar{q}_Z) \quad (44)$$

$$= -\int dZ q(Z) \log \frac{p(Y, Z|X, \theta, \pi) e^{-\sum_{k=1}^K G_k(Z, \bar{q})}}{q(Z)}, \quad (45)$$

$$G_k(Z, \bar{q}) = g(\bar{N}_k) + g'(\bar{N}_k)(N_k - \bar{N}_k), \quad (46)$$

$$N_k = \sum_{n=1}^N z_{nk}, \quad (47)$$

$$\bar{N}_k = \langle N_k \rangle_{\bar{q}(Z)} = \langle z_{nk} \rangle_{\bar{q}(Z)}, \quad (48)$$

$$g_{\text{AIC}}(N_k) = \frac{N_k}{N_k + \eta} (N + D_k), \quad (49)$$

$$g_{\text{BIC}}(N_k) = \frac{N_k}{2(N_k + \eta)} (N \log N + D_k \log N_k), \quad (50)$$

$$g'_{\text{AIC}}(N_k) = \frac{\eta}{(N_k + \eta)^2} (N + D_k), \quad (51)$$

$$g'_{\text{BIC}}(N_k) = \frac{\eta}{2(N_k + \eta)^2} \left(N \log N + D_k \log N_k + \frac{D_k}{\eta} (N_k + \eta) \right), \quad (52)$$

where we set $g = g_{\text{AIC}}$ and $g' = g'_{\text{AIC}}$ ($g = g_{\text{BIC}}$ and $g' = g'_{\text{BIC}}$) when we minimize AIC (BIC).

Using the upper bound in Eq. (45), we define the model selection problem as the following minimization problem:

$$\min_{\theta, \pi, q, \bar{q}} F(\theta, \pi, q_Z, \bar{q}_Z). \quad (53)$$

By minimizing $F(\theta, \pi, q_Z, \bar{q}_Z)$ in Eq. (45), we approximately minimize $E(Z, \theta)$.

We propose an algorithm for solving the problem in Eq. (53). We designate the algorithm as SICM algorithm because it minimizes the information criterion sequentially and approximately as in the variable selection case.

We summarize the SICM algorithm for model selection as Algorithm 2. The algorithm consists of the following four steps. First, we minimize F with respect to θ by solving the following maximum likelihood estimation problem:

$$\theta = \arg \min_{\theta} -\int dZ q(Z) \log p(Y, Z|X, \theta, \pi), \quad (54)$$

$$\theta_k = \arg \min_{\theta_k} -\sum_n r_{nk} \log p_k(y_n | \mathbf{x}_n, \theta_k) \quad (r_{nk} = \langle z_{nk} \rangle_{q(Z)}). \quad (55)$$

Second, F is minimized with respect to π by maximum likelihood estimation:

$$\pi_k = \frac{1}{N} \sum_n \langle z_{nk} \rangle_{q(Z)} = \frac{1}{N} \sum_n r_{nk}. \quad (56)$$

Third, F is minimized with respect to $q(Z)$. We assume that $q(Z)$ factorizes so that

$$q(Z) = \prod_{n=1}^N q_n(z_n), \quad (57)$$

$$\mathbf{z}_n = (z_{n1}, z_{n2}, \dots, z_{nK})^T. \quad (58)$$

Then, using the variational inference method explained in Appendix B, $q(Z)$ is optimized as

$$q_n(\mathbf{z}_n) = \prod_{k=1}^K (r_{nk})^{z_{nk}}, \quad (59)$$

$$r_{nk} = \frac{\rho_{nk}}{\sum_{k'=1}^K \rho_{nk'}}, \quad (60)$$

$$\rho_{nk} = \pi_k p_k(y_n | \mathbf{x}_n, \theta_k) \exp[-g'(\bar{N}_k)], \quad (61)$$

$$\langle z_{nk} \rangle_{q(Z)} = r_{nk}. \quad (62)$$

Fourth, F is minimized with respect to $\bar{q}(Z)$. The solution of this problem is expressed as

$$\bar{q} = q. \quad (63)$$

This is true because (1) $\bar{q}_Z$ appears in $F(\theta, \pi, q_Z, \bar{q}_Z)$ with the form of $\langle G_k(Z, \bar{q}) \rangle_q$; (2) $G_k(Z, \bar{q})$ is represented as Eq. (46); (3) Inequality (41) holds for $g(N_k)$, which is a convex upward function with respect to N_k .

Algorithm 2 SICM for model selection.

---

1. Setting. Choose AIC or BIC as the information criterion to be minimized. Input the user-defined parameters: η (positive constant to approximate an L_0 norm), ϵ (threshold for convergence condition used in Step 8 below) and K (maximum number of mixtures).
2. Initialization. Input the training observations, (Y, X) . Represent an update step as s and set $s = 0$. Randomly determine initial values of $\{r_{nk}^{(s=0)}\}$ ($\sum_k r_{nk} = 1, r_{nk} \geq 0$). Conduct maximum likelihood estimation and define the estimators as $\theta^{(s=0)} = \theta_0^{(s=0)}$ and $\pi^{(s=0)}$. Estimate the objective function: $F^{(s=0)} = F(\theta^{(s=0)}, \pi^{(s=0)}, q^{(s+1)}, \bar{q}^{(s+1)})$.
3. Update distribution $q(Z)$:

$$\bar{N}_k = \sum_n r_{nk}^{(s)}, \quad (64)$$

$$\rho_{nk} = \pi_k^{(s)} p_k(y_n | \mathbf{x}_n, \theta_k^{(s)}) \exp[-g'(\bar{N}_k)], \quad (65)$$

$$r_{nk}^{(s+1)} = \frac{\rho_{nk}}{\sum_{k'} \rho_{nk'}}. \quad (66)$$

4. Update distribution $\bar{q}(Z)$ as shown below.

$$\bar{q}^{(s+1)} = q^{(s+1)}. \quad (67)$$

5. Update parameters θ by maximum likelihood estimation as presented below.

$$\theta_k^{(s+1)} = \underset{\theta_k}{\operatorname{argmin}} - \sum_n r_{nk}^{(s+1)} \log p_k(y_n | \mathbf{x}_n, \theta_k) \quad (r_{nk} = \langle z_{nk} \rangle_{q(Z)}). \quad (68)$$

6. Update mixing coefficients π by maximum likelihood estimation.

$$\pi_k^{(s+1)} = \frac{1}{N} \sum_n r_{nk}^{(s+1)}. \quad (69)$$

7. Update the objective function. Estimate the objective function as shown below.

$$F^{(s+1)} = F(\theta^{(s+1)}, \pi^{(s+1)}, q^{(s+1)}, \bar{q}^{(s+1)}). \quad (70)$$

8. Convergence determination. If the following convergence condition

$$F^{(s)} - F^{(s+1)} < \epsilon \quad (71)$$

is satisfied, then go to Step 9. Otherwise, set $s = s + 1$ and go back to Step 3.

9. Output the results and stop the algorithm.
-

SICM for Hierarchical Relevance Determination

In this section, we propose the SICM algorithm for solving the HRD problem. We construct the algorithm based on the SICM algorithms described above, i.e., those for variable selection and model selection, which are summarized as Algorithms 1 and 2.

Let us start from the SICM for model selection. When we consider the HRD task, it is necessary to treat $\{||\theta_k||_0\}$ as variables, whereas they are constants in the model selection case. Therefore, when considering the HRD problem, it is

necessary to make the following alterations to the SICM for model selection:

$$D_k \rightarrow ||\theta_k||_0, \quad (72)$$

$$G_k(Z, \bar{q}) \rightarrow G_k(Z, \bar{q}, \theta_k). \quad (73)$$

By these changes, the upper bound to be minimized for solving the HRD problem is derived as

$$F(\theta, \pi, q_Z, \bar{q}_Z) = - \int dZ q(Z) \log \frac{p(Y, Z|X, \theta, \pi) e^{-\sum_{k=1}^K G_k(Z, \bar{q}, \theta_k)}}{q(Z)}. \quad (74)$$

Under the alterations, the individual steps in Algorithm 2 are changed as follows. First, the estimation step of q is invariant except for the expression of $g'(\bar{N}_k)$. It is necessary to replace D_k included in $g'(\bar{N}_k)$ with $\|\theta_k\|_0$. Second, the estimation step of $\bar{q}$ is invariant because (1) the optimal $\bar{q}$ is expressed as $\bar{q} = q$, as discussed above, and (2) this result is independent of the expression of D_k (independent of the described above alterations). Third, the estimation step of π is invariant because the estimation of π is only dependent on $\{\langle z_{nk} \rangle_q\}$ and is independent of D_k . Fourth, the estimation step of θ_k changes because $\sum_{k=1}^K G_k(Z, \bar{q}, \theta_k)$ becomes dependent on θ_k . Because of the dependence, the estimation step of θ in the HRD case is expressed as

$$\begin{aligned} \theta &= \arg \min_{\theta} - \int dZ q(Z) \log \frac{p(Y, Z|X, \theta, \pi) e^{-\sum_{k=1}^K G_k(Z, \bar{q}, \theta_k)}}{q(Z)} \\ &= \arg \min_{\theta} \int dZ q(Z) \left[-\log p(Y, Z|X, \theta, \pi) + \sum_{k=1}^K G_k(Z, \bar{q}, \theta_k) \right] \\ &= \arg \min_{\theta} \sum_{k=1}^K F_k, \end{aligned} \quad (75)$$

$$\begin{aligned} F_k &= \langle - \sum_n z_{nk} \log \pi_k - \sum_n z_{nk} \log p_k(y_n | \mathbf{x}_n, \theta_k) \\ &\quad + g(\bar{N}_k) + g'(\bar{N}_k)(N_k - \bar{N}_k) \rangle_{q(Z)} \\ &= - \sum_n r_{nk} \log p_k(y_n | \mathbf{x}_n, \theta_k) + g(\bar{N}_k, D_k = \|\theta_k\|_0) \\ &\quad + \text{const.} \quad (\bar{q} = q). \end{aligned} \quad (76)$$

Therefore θ_k is estimated as presented below:

$$\theta_k = \arg \min_{\theta_k} F_k = \arg \min_{\theta_k} - \sum_n r_{nk} \log p_k(y_n | \mathbf{x}_n, \theta_k) + h(\theta_k), \quad (77)$$

$$h_{\text{AIC}}(\theta_k) = \frac{\bar{N}_k}{\bar{N}_k + \eta} \|\theta_k\|_0, \quad (78)$$

$$h_{\text{BIC}}(\theta_k) = \frac{1}{2} \cdot \frac{\bar{N}_k}{\bar{N}_k + \eta} \log \bar{N}_k \cdot \|\theta_k\|_0, \quad (79)$$

$$\bar{N}_k = \sum_{n=1}^N \langle z_{nk} \rangle_{\bar{q}(Z)} = \sum_{n=1}^N \langle z_{nk} \rangle_{q(Z)} = \sum_n r_{nk}, \quad (80)$$

where we set $h = h_{\text{AIC}}$ ($h = h_{\text{BIC}}$) when we minimize AIC (BIC).

If we make the following alterations, then the upper bound F in Eq. (26) in the variable selection case coincides with F_k in Eq. (77) in the HRD case, except for their constant terms:

$$c_{\text{IC}} \rightarrow \frac{\bar{N}_k}{\bar{N}_k + \eta} c_{\text{IC}}, \quad (81)$$

$$\begin{aligned} \log p(Y|X, \theta) &= \sum_n \log p(y_n | \mathbf{x}_n, \theta) \\ &\rightarrow \sum_n r_{nk} \log p_k(y_n | \mathbf{x}_n, \theta_k). \end{aligned} \quad (82)$$

Therefore we can estimate θ_k using the SICM for variable selection (Algorithm 1) with the alterations in Eqs. (81) and (82). As a result, the problem for estimating θ_k becomes a weighted maximum likelihood estimation with L_1 regularization. The minimization of F_k is interpreted as a minimization of the information criterion, where the pdf is $p_k(y | \mathbf{x}, \theta_k)$ and the number of observations is N_k because (1) $r_{nk} \log p_k(y_n | \mathbf{x}_n, \theta_k)$ is considered to be a weighted log-likelihood of N_k observations because $0 \leq r_{nk} \leq 1$ and $\sum_n r_{nk} = N_k$; (2) $\frac{\bar{N}_k}{\bar{N}_k + \eta} c_{\text{IC}}$ is approximately equal to c_{IC} when N_k is larger than η , while we can adopt small η and we need not estimate θ_k if $N_k = 0$.

Summarizing the points presented above, the SICM algorithm can be constructed for solving the HRD problem in the following way. First, we replace Step 5 of Algorithm 2 (SICM for model selection) with Algorithm 1 (SICM for variable selection). Second, we make the alterations in Eqs. (81) and (82) to Algorithm 1. As an example, Appendix D summarizes the estimation step of θ_k , which corresponds to Step 5 in Algorithm 2 (corresponds to Algorithm 1 with the alterations), for the HRD problem of a logistic regression mixture. In Algorithm 3, we summarize the overall flow of the SICM algorithm for solving the HRD problem. The objective function F in Algorithm 3 is represented as

$$\begin{aligned} F &= - \sum_{n,k} r_{nk} \log p_k(y_n | \mathbf{x}_n, \theta_k) + \sum_k g(\bar{N}_k) \big|_{D_k = \|\theta_k\|_0} \\ &\quad + \sum_{n,k} r_{nk} \log \frac{r_{nk}}{\pi_k}, \end{aligned} \quad (83)$$

which is derived from Eqs. (49), (50), (74), (75) and (76).

Algorithm 3 Overall flow of the SICM algorithm for solving the HRD problem.

Require: training dataset for binary classification, type of information criterion to be minimized (AIC or BIC), maximum number of components K , L_0 approximation constant η , and convergence threshold ϵ .

Ensure: model assignment weights $\{r_{nk}\}_{n=1,k=1}^{n=N,k=K}$, mixing coefficients $\{\pi_k\}_{k=1}^K$, and model parameters of the individual mixtures $\{\theta_k\}_{k=1}^K$.

Initialization:

initialize the assignment weights $\{r_{nk}\}$ ($0 \leq r_{nk} \leq 1$, $\sum_{k=1}^K r_{nk} = 1$), and derive the mixing coefficients $\{\pi_k\}_{k=1}^K$ using Eq. (69).

Information criterion minimization:

continuously minimize the information criterion:

while the objective function F is not converged **do**

component parameter optimization: update $\{\theta_k\}_{k=1}^K$ with fixed $\{r_{nk}\}_{n=1,k=1}^{n=N,k=K}$ and fixed $\{\pi_k\}_{k=1}^K$. As an example, Appendix D and Algorithm 4 summarize the estimation step of θ_k in the case of logistic regression mixtures. This optimization step corresponds to variable selection.

assignment optimization: update $\{r_{nk}\}_{n=1,k=1}^{n=N,k=K}$ and $\{\pi_k\}_{k=1}^K$ with fixed $\{\theta_k\}_{k=1}^K$. This optimization step is conducted using Eqs. (64)–(66) and (69), and the step corresponds to model selection.

end while

return $\{\theta_k\}_{k=1}^K$, $\{r_{nk}\}_{n=1,k=1}^{n=N,k=K}$ and $\{\pi_k\}_{k=1}^K$.

Properties of SICM

In this section, we describe some properties of the SICM algorithm for HRD.

Sparsity

The SICM algorithm derives a sparse solution theoretically as discussed below.

Automatically, SICM determines the number of mixture components. In fact, Eq. (65) in Algorithm 2 plays a fundamentally important role for the determination. By the effect of the term $\exp[-g'(\tilde{N}_k)]$, a mixture component with a few observations is erased. Thereby, the number of the components is optimized because $\exp[-g'(\tilde{N}_k)]$ becomes close to 0 if $\tilde{N}_k$, representing the number of observations of the k -th component, becomes small. Consequently, ρ_{nk} rapidly converges to 0 if $\tilde{N}_k$ becomes smaller.

The parameters of the individual components are made sparse by SICM. As discussed in “[SICM for Hierarchical Relevance Determination](#)”, SICM estimates the parameters of the individual mixtures by solving the L_1 -regularized maximum likelihood estimation problems. Therefore SICM derives sparse parameters of the individual components by the regularization effect of L_1 penalty, as in the case of sparse estimation methods such as Lasso.

Monotonicity

The SICM algorithm for HRD monotonically decreases the objective function, which corresponds to the upper bound

of the information criterion to be minimized. We present a sketch of the proof below.

SICM for model selection monotonically decreases F in Eq. (45), the upper bound of the information criterion to be minimized. A sketch of the proof can be shown as

$$\begin{aligned} F^{(s)} &= F(\theta^{(s)}, \pi^{(s)}, q^{(s)}, \bar{q}^{(s)}) \\ &\geq F(\theta^{(s)}, \pi^{(s)}, q^{(s+1)}, \bar{q}^{(s)}) \end{aligned} \quad (84)$$

$$\geq F(\theta^{(s)}, \pi^{(s)}, q^{(s+1)}, \bar{q}^{(s+1)}) \quad (85)$$

$$\geq F(\theta^{(s+1)}, \pi^{(s+1)}, q^{(s+1)}, \bar{q}^{(s+1)}), \quad (86)$$

where s represents an update step. Inequality (84) arises from Eqs. (90) and (92) in Appendix B, and Step 3 of Algorithm 2, which estimates the optimal q variationally by fixing $\bar{q}$, θ and π . Inequality (85) arises from Step 4 of Algorithm 2, which estimates the optimal $\bar{q}$ [as described immediately after Eq. (62)] by fixing q , θ and π . Inequality (86) arises from Step 5 and Step 6, which conducts maximum likelihood estimations of θ and π by fixing q and $\bar{q}$.

In the case of SICM for HRD, which is a combination of SICM for model selection and SICM for variable selection, the maximum likelihood estimation of θ is replaced with L_1 -penalized sparse estimation of θ such as sparse logistic regression summarized in Appendix D. The sparse estimation decreases the upper bound expressed as the L_1 -regularized objective function. Therefore inequality (86) also holds in the HRD case.

Summarizing the explanation presented above, the SICM for HRD satisfies inequalities (84)–(86). It therefore monotonically decreases the objective function.

Fig. 1 Example of the interpretable hierarchical model based on the SICM algorithm. In accordance with the rules represented as the decision tree, the prediction model used for an observation is switched. In the figure, squares and circles respectively represent conditional branches and assigned regression models. $\{A, B, C\}$ and $\{\bar{A}, \bar{B}, \bar{C}\}$ respectively represent conditions and their complements. The number of regression models used in the hierarchical model and the subsets of input variables used in the individual regressions are determined automatically by the SICM algorithm

Properties Related to Information Criterion Minimization

The SICM algorithm can avoid the singularity problem of mixture modeling described hereinafter. Both AIC and BIC are derived based on second-order expansions of their original objective functions and are derived by assuming the regularity condition. It is therefore not justifiable to apply the information criteria to singular models. However, a mixture model represented as $p(y|\mathbf{x}, \theta) = \sum_k \pi_k p_k(y|\mathbf{x}, \theta_k)$ is singular. SICM avoids this difficulty for the following reasons. When constructing SICM, we started from a mixture represented as Eq. (2) and considered the objective function corresponding to its information criterion in Eq. (19). $p(Y|X, Z, \theta)$ in Eq. (2) is regular because it is assumed that the conditions described in “Hierarchical Relevance Determination Problem” (“each mixture $p_k(y|\mathbf{x}, \theta)$ is regular” and “the mixture assignment of each observation is unique”) hold. Consequently, use of the information criterion corresponding to Eq. (19) is justified.

Actually, SICM with BIC minimization has asymptotic consistency. Because $p(Y|X, Z, \theta)$ satisfies the regularity condition as described above, its Laplace approximation (BIC used in SICM) has asymptotic consistency. Actually, SICM with AIC does not have asymptotic consistency as AIC does not. Nevertheless, AIC is “consistent” in the sense that an estimated distribution asymptotically approaches the true distribution (not the true “model” in the case of BIC).

One benefit of SICM is that it can minimize either AIC or BIC, whereas Bayesian methods such as variational Bayesian inference correspond only to BIC minimization (actually, minimization of negative marginal log-likelihood). We

expected, and we show in “Experiments”, that SICM with AIC sometimes outperforms SICM with BIC in some cases.

Applicability to Unsupervised Learning

As described in this report, we have considered the task of relevance determination in supervised mixture distributions. However, the SICM algorithm is applicable to unsupervised mixtures by making the following alteration:

$$p_k(y_n|\mathbf{x}_n, \theta_k) \rightarrow p_k(\mathbf{x}_n|\theta_k). \quad (87)$$

Unsupervised mixtures include a mixture of normal distributions. For instance, by application of SICM to a normal mixture, we can estimate a mixture of sparse Gaussian graphical models.

Application to Interpretable Hierarchical Modeling

In this section, as an application of SICM, we propose a method for constructing an interpretable hierarchical model, which is constructed as a combination of interpretable models: a decision tree and regression models.

The overall flow of the model construction is summarized as presented below.

1. We estimate $p(Y|X, Z, \theta)$ using the SICM algorithm. As the individual mixtures, we use interpretable regression models: linear regression models or logistic regression models.
2. We construct a decision tree for model assignment using the training dataset where $\{r_{nk}\}_{k=1}^K$ ($r_{nk} = \langle z_{nk} \rangle_q$) is a set of target variables and where $\mathbf{x}_n$ is a set of input variables. The decision tree for predicting $r_{\cdot k}$ is estimated by solving a K -class classification problem.
3. Using the results of the two steps described above, we conduct prediction as follows:

$$p(y|\mathbf{x}) = \sum_k r_k(\mathbf{x}) p_k(y|\mathbf{x}, \theta_k). \quad (88)$$

Here $r_k(\mathbf{x})$ represents the predicted value of $r_{\cdot k}$ estimated from a test observation $\mathbf{x}$ using the decision tree. Actually, $r_k(\mathbf{x})$ is considered $p(k|\mathbf{x})$.

A key point in constructing the model is the introduction of Step 2, the decision tree estimation. To conduct prediction, it is necessary to estimate the predicted value of $r_{\cdot k}$ from a test observation. However, the prediction mechanism of $r_{\cdot k}$ is not included in the SICM algorithm. Therefore, Step 2 is introduced to predict $r_{\cdot k}$. Consequently, we adopt the two-step training process described above, which consists of

Table 1 Properties of nine UCI datasets used in the experiments

Dataset name	# of records	# of numerical input variables	# of categorical input variables
Bank marketing	41187	10	10
Breast cancer Coimbra	116	9	1
Climate model simulation crashes	540	20	0
Default of credit card clients	30000	20	3
Diabetic retinopathy debrecen	1151	19	0
Indian liver patients	583	10	1
QSAR biodegradation	1055	41	0
Seismic bumps	2584	14	4
Spambase	4601	57	0

The numbers of input variables do not include variables not used for model construction (ex. ID variables and time stamps)

HRD based on SICM and model assignment determination based on a decision tree.

The hierarchical model is constructed as a combination of a decision tree and regression models (linear regression models or logistic regression models). When conducting prediction according to the hierarchical model, the regression model used for an observation is switched in accordance with the rules in the decision tree. Figure 1 presents an example of the interpretable hierarchical models. For example, in the case of Fig. 1, the predicted value is estimated (1) by Model (regression model) 1 ($p(y|\mathbf{x}) = p_1(y|\mathbf{x}, \theta_1)$) if the observation $\mathbf{x}$ satisfies Condition A; (2) by Model 2 if $\mathbf{x}$ satisfies Conditions $\bar{A}$, B and C ; (3) by Model 3 if $\mathbf{x}$ satisfies Conditions $\bar{A}$, B and $\bar{C}$; and (4) by Model 4 if $\mathbf{x}$ satisfies Conditions $\bar{A}$ and $\bar{B}$. The proposed model differs from ensemble learning models because a single model is assigned to an observation in the case of the proposed model. When estimating the model, the number of regression models used in the hierarchical model and the subsets of input variables used in the individual regression models are determined automatically using the SICM algorithm.

Experiments

Experimental Setting

Through the following experiments, we demonstrate the utility of the SICM algorithm using the hierarchical prediction model proposed in “Application to Interpretable Hierarchical Modeling”.

We considered binary classification problems in the experiments. Therefore, we used a logistic regression as a mixture component of the hierarchical model. Appendix D summarizes the SICM algorithm for logistic regression mixtures. The maximum number of components, the constant for approximating the L_0 terms, and

the threshold for the convergence condition were set to $(K, \eta, \epsilon) = (50, 1.0, 10^{-6})$. The maximum number of components K is equal to the initial number of components in the SICM algorithm. The number of components decreases as the SICM steps progress, but it never increases. Therefore it is necessary to set K to enough large number and we set K to 50. The approximation constant η was not estimated from training datasets but was given by us as a fixed value.

We used nine datasets from the UCI repository [19]. We selected these datasets from the UCI repository because they are datasets for binary classification problems and consist of a binary target variable and numerical input variables. Some properties of the datasets are summarized in Table 1.

For training the models, we transformed categorical input variables into their one-hot encodings and standardized all input variables.

We observed the binary classification accuracies of the proposed models: SICM_{AIC} and SICM_{BIC} . Hereinafter, we denote the proposed hierarchical model estimated by minimizing AIC (BIC) as SICM_{AIC} (SICM_{BIC}). As a measure of the classification accuracy, we used the mean Area Under the Curve (AUC) of ROC curves estimated using five-fold cross validation.

In the experiments, we aimed to show the effectiveness of the SICM algorithm by demonstrating (a) we could construct accurate and interpretable models by solving the HRD problem, and (b) we could get better results using the SICM algorithm than using another existing inference method. In other words, (a) means that we could improve prediction accuracies of interpretable models by solving the HRD problem. In order to show (a), we compared SICMs (SICM_{AIC} and SICM_{BIC} , the hierarchical models proposed in “Application to Interpretable Hierarchical Modeling”, which are estimated using the SICM algorithm) to decision tree and logistic regression, which are frequently used as interpretable models. For this comparison, as our model, we selected the hierarchical prediction model proposed in “Application

Table 2 Classification accuracies of six prediction models

Datasets	SICM _{AIC}	SICM _{BIC}	VB	DT	LOGIT	SVM
Bank marketing	0.856	0.845	0.844	0.722	0.720	0.874
Breast cancer Coimbra	0.820	0.853	0.802	0.685	0.745	0.763
Climate model simulation crashes	0.974	0.973	0.962	0.691	0.868	0.907
Default of credit card clients	0.767	0.756	0.733	0.694	0.716	0.765
Diabetic retinopathy Debrecen	0.832	0.849	0.814	0.582	0.733	0.762
Indian liver patients	0.737	0.751	0.719	0.544	0.685	0.705
QSAR biodegradation	0.927	0.927	0.916	0.797	0.859	0.878
Seismic bumps	0.733	0.744	0.736	0.529	0.707	0.750
Spambase	0.982	0.974	0.965	0.914	0.931	0.936

Each value represents a mean AUC. The mean AUC was evaluated by conducting five-fold cross validation. The best results are presented in bold typeface. DT and LOGIT respectively represent decision tree and logistic regression

to Interpretable Hierarchical Modeling”, which consisted of tree and logistic regression. Namely to compare to tree and logistic regression, we constructed the hierarchical model consisting of a tree and logistic regressions. For reference, we compared SICMs to support vector machine (SVM), which is a highly accurate uninterpretable model. We note that SICM is not proposed for constructing highly accurate uninterpretable models but proposed for improving prediction accuracies of interpretable models. Therefore, in the experiments, we did NOT aim to show SICMs outperformed SVM because SVM is uninterpretable. SVM was selected to observe how closely SICMs performs to an uninterpretable model.

In order to show (b) SICMs derived better results than an existing inference method, we compared SICMs to VB. Here, VB represents the hierarchical prediction model which has the same structure as SICMs and its parameter inference method is replaced from the SICM algorithm to variational Bayesian inference (namely SICMs and VB are the same models estimated by the different methods). We selected VB for the following reasons. First, as described in “Introduction”, our main aim and contribution are to construct a new inference method for solving the HRD task. Therefore, we compared the SICM algorithm to variational Bayesian inference by comparing the same models estimated by different

inference methods. Variational Bayesian inference is one of the most popular method for relevance determination and is still a SOTA in this area. Second, in this paper, we do NOT aim to propose a new hierarchical model, and thus to select an optimal model family is out of scope. Therefore we selected VB having the same structure as SICMs, and did not select another latest hierarchical models belonging to different model families.

Let us give supplementary explanations to how to estimate the compared models. Logistic regression models were constructed by maximum likelihood estimation. Gaussian kernels were used in SVM. We optimized a decay parameter of a Gaussian kernel by conducting two-fold cross validation. As mentioned above, we used variational Bayesian inference for estimating VB.

The computational environment was as follows: Intel Xeon CPU ES-1650 3.50 GHz CPU and 64 GB memory on a Linux Ubuntu platform. We conducted the experiments using Python language. The implementations of SVM, logistic regression and decision tree (including the tree estimation part in VB and SICMs) were taken from scikit-learn [21], a public machine learning library in Python. We implemented the HRD part of VB (variational Bayesian inference of a logistic regression mixture) based on Refs. [1, 22].

Results and Discussion

Prediction Performance

Table 2 presents the experimentally obtained results: the classification accuracies of six models. Results indicate the SICM algorithm’s effectiveness for the following reasons.

By combining decision tree and logistic regression, we can construct a model with higher accuracy than that of any individual models. We can construct an interpretable and highly accurate model using the SICM algorithm. This statement is supported by the following results. First,

Table 3 Model assignment rules of the proposed hierarchical model

Model	Model assignment rule
1	(bill statement in July $\leq -0.5\sigma$) and (payment in August $\leq 4.0\sigma$) and (payment delay in September $\leq 0.5\sigma$)
2	(bill statement in July $> -0.5\sigma$) and (payment in August $\leq 4.0\sigma$) and (payment delay in September $\leq 0.5\sigma$)
3	(payment in August $\leq 4.0\sigma$) and (payment delay in September $> 0.5\sigma$)
4	(payment in August $> 4.0\sigma$) and (payment delay in September $\leq 0.5\sigma$) and (payment in September $\leq -0.1\sigma$)
5	(payment in August $> 4.0\sigma$) and (payment delay in September $> 0.5\sigma$) and (payment in September $\leq -0.1\sigma$)
6	($4.0\sigma < \text{payment in August} \leq 6.9\sigma$) and (payment in September $> -0.1\sigma$)
4	(payment in August $> 6.9\sigma$) and (payment in September $> -0.1\sigma$)

Each row represents a model assignment rule, which corresponds to a leaf of a decision tree and which is expressed as a product set of the listed conditions. For example, for the first rule in the figure, we use Model 1 for prediction if an observation simultaneously satisfies the following conditions: (1) bill statement in July $\leq -0.5\sigma$; (2) payment in August $\leq 4.0\sigma$; and (3) payment delay in September $\leq 0.5\sigma$. In the rules, σ represent standard deviations of the respective input variables. The IDs of the assigned regression model (from Model 1 to Model 6) correspond to those in Table 4

Table 4 Regression coefficients of the individual logistic regression models included in the proposed hierarchical model

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
intercept	-3.37	-4.18	-2.98	-2.12	-4.27	-4.42
amount of the given credit	0.00	0.00	-1.94	-0.41	-0.14	0.00
age	0.00	-0.16	0.00	0.77	0.00	0.22
payment delay in September	-1.19	1.12	1.08	0.00	4.05	3.82
payment delay in August	0.35	-1.06	1.03	0.00	0.00	0.00
payment delay in July	0.43	-2.20	1.90	-0.12	0.00	-0.22
payment delay in June	0.21	2.39	0.29	0.00	0.28	-0.34
payment delay in May	-0.29	1.83	0.57	0.00	0.00	0.00
payment delay in April	0.00	1.98	-0.33	0.00	-0.37	0.00
bill statement in September	-1.31	0.00	0.36	0.00	0.00	-0.48
bill statement in August	-1.16	0.00	0.96	1.72	0.80	0.00
bill statement in July	-2.15	0.00	0.71	0.00	-0.81	0.00
bill statement in June	0.00	-0.71	0.00	0.00	0.00	0.51
bill statement in May	-0.64	2.01	-1.19	0.00	0.00	0.00
bill statement in April	0.00	0.00	0.00	0.00	0.43	0.33
payment in September	0.00	0.00	-0.22	-14.24	-0.66	0.00
payment in August	0.13	-1.91	-1.72	1.92	0.45	-0.37
payment in July	-0.15	0.00	0.35	-4.70	-0.25	-0.14
payment in June	0.00	0.00	0.06	0.39	-0.03	-0.20
payment in May	-0.28	0.80	-0.13	-0.11	-0.68	-1.08
payment in April	-0.10	-0.71	0.44	0.14	-0.53	-0.17
male	0.00	0.00	0.31	0.38	0.39	0.52
graduate school	1.80	0.30	-0.40	0.60	0.00	2.10
university	1.76	0.00	0.00	0.74	0.18	2.13
high school	1.40	0.48	-0.12	0.00	0.00	1.71
married	1.78	0.75	0.00	0.00	0.00	0.48
single	1.52	0.36	0.00	0.00	0.00	0.53

Rows and columns respectively represent an input variable and a logistic regression model. One logistic regression model in this table was used for prediction in accordance with the rules in Table 3. A coefficient is presented with red/white/blue color if its value is positive/exactly zero/negative. If a coefficient is positive/negative, then the default risk increases when the corresponding input variable increases/decreases

combination models (hierarchical models), known as SICM_{AIC} , SICM_{BIC} and VB, outperformed decision tree and logistic regression. Second, the combination models performed comparably to SVM.

The SICM algorithm which minimizes BIC is theoretically better than variational Bayesian inference because the independence between the latent variables and the parameters of the individual components is not assumed in SICM, whereas variational Bayesian inference requires the

independence assumption. This statement is supported by the results that SICM_{BIC} outperformed VB.

One benefit of SICM is that SICM can minimize either AIC or BIC, whereas Bayesian methods such as variational Bayesian inference only minimize the negative marginal log-likelihood. In fact, BIC is an approximate representation of the negative marginal log-likelihood. This statement is supported by results showing that SICM_{AIC} performed comparably to SICM_{BIC} from the viewpoint of their win-loss record.

Interpretability

Using the proposed method, we can construct an interpretable model because the proposed hierarchical model is represented as a combination of a decision tree and regression models, and has the sparse structure. In this section, we qualitatively describe the property of interpretability through an example of the proposed model.²

As a related example, we used the hierarchical model trained on “Default of Credit Card Clients” dataset. An observation in the dataset represents a record of a credit card user, and consists of a binary target variable and 23 input variables. The target variable represents default on a payment in October. The input variables represent (1) amount of the given credit; (2) payment status recorded in April–September such as payment delay (month), amount of bill statement, and amount of payment; and (3) user profile such as age, gender, education, and marital status. We transformed the categorical input variables into their one-hot encodings.

Tables 3 and 4 summarize the proposed hierarchical model trained on the “default of credit card clients” dataset based on the SICM algorithm. Table 3 represents the model assignment rules included in the decision tree. Table 4 shows regression coefficients of the individual logistic regression models. These results indicate that the proposed model has the following properties.

The SICM algorithm derives sparse solutions for the following reasons. First, as shown in Table 4, many coefficients were estimated as exactly zero. Second, the number of regression models diminished to six starting with $K = 50$.

The proposed model assigns a regression model to an observation based on its properties. For example, in the case of Table 3, (1) Model 1 corresponded to the “standard” clients (observations), who had low payment delay in September and low bill statement in July; (2) Model 2 corresponded to the clients who had larger bill statements in July than the standard clients had; (3) Model 3 corresponded to the clients who had more months of payment delay in September than the standard clients; and (4) Models 4, 5, and 6 corresponded to clients who had exceptionally large payments in August. Each regression model occupied an imbalanced number of

observations. Three models (Models 4, 5 and 6) were used for the “exceptional” clients.

For each input variable, the signs of the regression coefficients may differ depending on the regression model. For example, in the case of Model 2, the regression coefficient of “age” was negative, which indicates that the default risk decreases as age increases. However, for Models 4 and 6, the coefficients were positive, which indicates that the default risk increases when age increases. Age inversely contributed to the default risk depending on whether the payment in August was exceptionally large ($> 4.0\sigma$), or not. Such local sign inversion is regarded as one reason why the proposed hierarchical model is highly accurate.

Sign inversion does not always occur. For example, “the amount of the given credit” invariably had non-positive coefficients, which indicates that the default risk always decreases as the credit rating becomes better. This result is consistent with our intuition.

Summary

The SICM algorithm was proposed for solving the hierarchical relevance determination problem. The SICM algorithm minimizes an information criterion continuously and therefore enables us to determine the degrees of freedom in a mixture distribution automatically. A method for constructing an interpretable hierarchical model based on the SICM algorithm was also proposed. Experiment results obtained using the interpretable hierarchical model have demonstrated the utility of the SICM algorithm for the following reasons. First, the hierarchical model outperformed frequently used interpretable models (tree and logistic regression) in terms of prediction accuracy. Second, it was shown qualitatively that the hierarchical model derived interpretable results consistent with our intuition.

Future work includes the following two issues. One is a theoretical expansion of the interpretable hierarchical model. When constructing the model, the degrees of freedom in the decision tree for model assignment are not automatically determined. Therefore introduction of relevance determination mechanism to the decision tree estimation is a subject of future work. The other is an application of the proposed information criterion minimization method to those other than the HRD problem. Relevance determination based on the continuous minimization of the information criterion is widely applicable. A promising application is relevance determination in unsupervised distributions such as sparse estimation of Gaussian graphical model mixtures.

² Here we did not compare interpretability of the models for the following reasons though we compared the prediction accuracies of them. In machine learning context, *interpretable* means that a model satisfies the following conditions: (1) how to derive the output value (ex. probability of $y = 1$) is analytically trackable; (2) it is trackable which variables affect a model and how much. By the definition, we can only observe whether a model is interpretable or uninterpretable, and thus it is difficult to measure how much the model is interpretable. In the case of the experiments, SVM is not interpretable (therefore is less interpretable than SICMs), and decision tree, logistic regression and SICMs are interpretable (therefore they are equally interpretable).

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

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/>.

Appendix A: Jensen's Inequality

Jensen's inequality is expressed as

$$f\left(\int d\mathbf{x} q(\mathbf{x})g(\mathbf{x})\right) \geq \int d\mathbf{x} q(\mathbf{x})f(g(\mathbf{x})), \quad (89)$$

where f , g , and q respectively represent a convex upward function, an arbitrary function, and a probability density function (pdf).

Appendix B: Minimization Problem of KL-Divergence-Like Functions

Here let us explain the variational minimization of KL-divergence-like functions. We consider the following minimization problem of

$$\min_q - \int d\mathbf{x} q(\mathbf{x}) \log \frac{H(\mathbf{x})}{q(\mathbf{x})} \quad (H(\mathbf{x}) > 0), \quad (90)$$

where q and H respectively represent a pdf and a positive function. If q factorizes such that

$$q(\mathbf{x}) = \prod_{i=1}^D q_i(\mathbf{x}_i) = q_i(\mathbf{x}_i) q_{\bar{i}}(\mathbf{x}_{\bar{i}}) \left(q_{\bar{i}}(\mathbf{x}_{\bar{i}}) = \prod_{j \neq i}^D q_j(\mathbf{x}_j) \right), \quad (91)$$

then the problem has the following solution [23]:

$$\log q_i(\mathbf{x}_i) = \int d\mathbf{x}_{\bar{i}} q_{\bar{i}}(\mathbf{x}_{\bar{i}}) \log H(\mathbf{x}) + \text{const.}, \quad (92)$$

where i and $\bar{i}$ respectively represent a set of variables and its complementary set.

Appendix C: L_1 -regularized Quadratic Programming Problem

Let us explain the sequential method for minimizing an L_1 -regularized quadratic function. We consider the following minimization problem:

$$\min_{\mathbf{x}} E, \quad (93)$$

$$E(\mathbf{x}) = \frac{1}{2} \mathbf{x}^\dagger W \mathbf{x} - \mathbf{s}^\dagger \mathbf{x} + \|\mathbf{a} \circ \mathbf{x}\|_1, \quad (94)$$

where $\circ$ represents Hadamard product. In Eq. (94), $\mathbf{x}$, $\mathbf{s}$, and $\mathbf{a}$ represent D -dimensional vectors, and W represents a $D \times D$ matrix. We assume that the following conditions are satisfied:

$$a_i > 0, \quad (95)$$

$$W_{ii} > 0, \quad (96)$$

$$W \geq 0. \quad (97)$$

The problem is solvable with a coordinate-wise subgradient method [24]. The algorithm is sketched below. Using a subderivative, let us define a function $d(\cdot)$ as

$$d(z) = \frac{\partial}{\partial z} |z| = \begin{cases} +1 & (z > 0) \\ [-1, +1] & (z = 0) \\ -1 & (z < 0), \end{cases} \quad (98)$$

which represents the sign of the input scalar z . The optimal solution of the problem is derived by solving $\frac{\partial}{\partial \mathbf{x}} E = \mathbf{0}$ with respect to $\mathbf{x}$. The i th component of this equation is represented as

$$\frac{\partial}{\partial x_i} E = \sum_j W_{ij} x_j - s_i + a_i d(x_i) = W_{ii} x_i - u_i + a_i d(x_i) = 0, \quad (99)$$

$$u_i = s_i - \sum_{j \neq i} W_{ij} x_j. \quad (100)$$

By solving Eq. (99) with respect to x_i , one obtains

$$x_i = \frac{1}{W_{ii}} [u_i - a_i d(x_i)]. \quad (101)$$

Considering the sign of Eq. (101), by substituting Eq. (101) to a sign function, we get

$$\begin{aligned} d(x_i) &= d\left(\frac{1}{W_{ii}} [u_i - a_i d(x_i)]\right) = d\left(\frac{1}{W_{ii}}\right) d(u_i - a_i d(x_i)) \\ &= d(u_i - a_i d(x_i)). \end{aligned} \quad (102)$$

Equation (102) is derived because $W_{ii} > 0$ holds. From Eq. (102), the following conditions for relations between u_i and a_i are derived:

$$d(u_i + a_i) = -1 \quad \text{for } x_i < 0, \quad (103)$$

$$d(u_i - a_i d(x_i)) = d(x_i) \quad \text{for } x_i = 0, \quad (104)$$

$$d(u_i - a_i) = 1 \quad \text{for } x_i > 0. \quad (105)$$

By combining Eqs. (101) and (103)–(105), the following solution is obtained:

$$x_i = \begin{cases} \frac{1}{W_{ii}}(u_i - a_i) & (u_i > a_i) \\ 0 & (-a_i \leq u_i \leq a_i) \\ \frac{1}{W_{ii}}(u_i + a_i) & (u_i < -a_i) \end{cases}. \quad (106)$$

Equation (106) is an update equation for the i th component. By repeating the update step until convergence, the solution of the L_1 -regularized quadratic programming problem can be derived.

Appendix D: SICM Algorithm for Hierarchical Relevance Determination in Logistic Regression Mixtures

We explain the method for estimating the parameters of the individual mixtures when we apply the SICM algorithm for HRD in a logistic regression mixture.

We consider the HRD problem of a logistic regression mixture, where each mixture component is expressed as

$$p_k(y = +1 | \mathbf{x}, \mathbf{w}_k) = \frac{1}{1 + \exp[-\mathbf{w}_k^\dagger \mathbf{x}]}, \quad (107)$$

$$\begin{aligned} p_k(y = -1 | \mathbf{x}, \mathbf{w}_k) &= \frac{\exp[-\mathbf{w}_k^\dagger \mathbf{x}]}{1 + \exp[-\mathbf{w}_k^\dagger \mathbf{x}]} \\ &= \frac{1}{1 + \exp[\mathbf{w}_k^\dagger \mathbf{x}]}, \end{aligned} \quad (108)$$

$$p_k(y | \mathbf{x}, \mathbf{w}_k) = \sigma(y \mathbf{w}_k^\dagger \mathbf{x}), \quad (109)$$

$$\sigma(z) = \frac{1}{1 + \exp[-z]}, \quad (110)$$

where σ represents a sigmoid function.

As discussed in “SICM for Hierarchical Relevance Determination”, the parameters of the k th component are estimated by minimizing the following function F_k as

$$F_k = IC(Y|X, Z_k, \mathbf{w}_k) = \mathcal{L}_k + J_{\text{IC}}(\bar{N}_k) \|\mathbf{w}_k\|_0, \quad (111)$$

$$\mathcal{L}_k = -\sum_n r_{nk} \log p(y_n | \mathbf{x}_n, \mathbf{w}_k), \quad (112)$$

$$J_{\text{AIC}}(\bar{N}_k) = \frac{\bar{N}_k}{\bar{N}_k + \eta}, \quad (113)$$

$$J_{\text{BIC}}(\bar{N}_k) = \frac{1}{2} \frac{\bar{N}_k}{\bar{N}_k + \eta} \log \bar{N}_k, \quad (114)$$

$$r_{nk} = \langle z_{nk} \rangle_{q(Z)}, \quad (115)$$

$$\bar{N}_k = \sum_n r_{nk}. \quad (116)$$

Sigmoid function σ has the following tight lower bound [25]:

$$\sigma(z) \geq \sigma(\xi) \exp \left[\frac{1}{2} (z - \xi) - \lambda(\xi)(z^2 - \xi^2) \right], \quad (117)$$

$$\lambda(\xi) = \frac{1}{2\xi} \left(\sigma(\xi) - \frac{1}{2} \right), \quad (118)$$

where ξ is an additional parameter per an observation. Applying this bound, the following upper bound of $\mathcal{L}_k$ is derived:

$$\begin{aligned} \mathcal{L}_k &\leq \mathcal{H}_k(\mathbf{w}_k, \xi_k) \\ &= -\sum_n r_{nk} \left[-\mathbf{w}_k^\dagger (\lambda(\xi_{kn}) \mathbf{x}_n \mathbf{x}_n^\dagger) \mathbf{w}_k^\dagger + \mathbf{w}_k^\dagger \left(\frac{1}{2} y_n \mathbf{x}_n \right) \right. \\ &\quad \left. + \left(\log \sigma(\xi_{kn}) - \frac{1}{2} \xi_{kn} + \lambda(\xi_{kn}) \xi_{kn}^2 \right) \right] \end{aligned} \quad (119)$$

$$= \frac{1}{2} \mathbf{w}_k^\dagger W_k \mathbf{w}_k - \mathbf{s}_k^\dagger \mathbf{w}_k + C_k, \quad (120)$$

$$W_k = 2 \sum_n r_{nk} \lambda(\xi_{kn}) \mathbf{x}_n \mathbf{x}_n^\dagger, \quad (121)$$

$$\mathbf{s}_k = \frac{1}{2} \sum_n r_{nk} y_n \mathbf{x}_n, \quad (122)$$

$$\begin{aligned} C_k &= C_k(\xi_k) \\ &= \sum_n r_{nk} \left(\frac{1}{2} \xi_{kn} - \log \sigma(\xi_{kn}) - \lambda(\xi_{kn}) \xi_{kn}^2 \right), \end{aligned} \quad (123)$$

where ξ_k is an N -dimensional parameter vector and ξ_{kn} represents its n th component. Applying this bound and the upper

bound of $\|\mathbf{w}_k\|_0$ in Eq. (24), the following upper bound of the objective function F_k is obtained.

$$\begin{aligned} F_k &= IC(Y|X, Z_k, \mathbf{w}_k) = \mathcal{L}_k + J_{IC}(\bar{N}_k) \|\mathbf{w}_k\|_0 \\ &\leq \mathcal{H}_k(\mathbf{w}_k, \xi_k) + J_{IC}(\bar{N}_k) \|\mathbf{w}_k\|_0 \\ &\leq \mathcal{F}_k(\mathbf{w}_k, \mathbf{w}_{k0}, \xi_k), \end{aligned} \quad (124)$$

$$\begin{aligned} \mathcal{F}_k(\mathbf{w}_k, \mathbf{w}_{k0}, \xi_k) &= \mathcal{H}_k(\mathbf{w}_k, \xi_k) \\ &\quad + J_{IC}(\bar{N}_k) \left[\sum_i \frac{\eta}{(|w_{k0i}| + \eta)^2} |w_{ki}| + \frac{|w_{k0i}|^2}{(|w_{k0i}| + \eta)^2} \right] \\ &= \frac{1}{2} \mathbf{w}_k^\dagger W_k \mathbf{w}_k - \mathbf{s}_k^\dagger \mathbf{w}_k + \|\mathbf{a}_k \circ \mathbf{w}_k\|_1 \\ &\quad + C_k(\xi_k) + J_{IC}(\bar{N}_k) \sum_i \frac{|w_{k0i}|^2}{(|w_{k0i}| + \eta)^2}, \end{aligned} \quad (125)$$

$$(\mathbf{a}_k)_i = J_{IC}(\bar{N}_k) \frac{\eta}{(|w_{k0i}| + \eta)^2}. \quad (126)$$

When deriving the upper bound, we introduced the additional parameters, $\mathbf{w}_{k0}$ and ξ_k . As discussed in “Proposed Method: SICM Algorithm”, F_k can be minimized approximately by minimizing its upper bound: $\mathcal{F}_k(\mathbf{w}_k, \mathbf{w}_{k0}, \xi_k)$.

We erase ξ_k from $\mathcal{F}_k(\mathbf{w}_k, \mathbf{w}_{k0}, \xi_k)$ in the following manner. When $\mathbf{w}_k$ and ξ_k are given, the optimal $\mathbf{w}_{k0}$, which minimizes $\mathcal{F}_k(\mathbf{w}_k, \mathbf{w}_{k0}, \xi_k)$, is expressed as

$$\mathbf{w}_{k0} = \mathbf{w}_k. \quad (127)$$

In inequality (117), the equality holds if $z = \xi$. Presuming $z = z_n = y_n \mathbf{w}_k^\dagger \mathbf{x}_n$ and $\xi = \xi_{kn}$ in the inequality, the optimal ξ_k , which minimizes $\mathcal{F}_k$ when $\mathbf{w}_k$ is given, is expressed as

$$\xi_{kn} = y_n \mathbf{w}_k^\dagger \mathbf{x}_n. \quad (128)$$

From Eqs. (127) and (128), we select ξ_{kn} as

$$\xi_{kn} = y_n \mathbf{w}_{k0}^\dagger \mathbf{x}_n, \quad (129)$$

and erase ξ_k from $\mathcal{F}_k$ as shown below:

$$\begin{aligned} F_k &= IC(Y|X, Z_k, \mathbf{w}_k) \\ &\leq \mathcal{F}_k(\mathbf{w}_k, \mathbf{w}_{k0}), \end{aligned} \quad (130)$$

$$\begin{aligned} \mathcal{F}_k(\mathbf{w}_k, \mathbf{w}_{k0}) &= \frac{1}{2} \mathbf{w}_k^\dagger W_k \mathbf{w}_k - \mathbf{s}_k^\dagger \mathbf{w}_k + \|\mathbf{a}_k \circ \mathbf{w}_k\|_1 \\ &\quad + C_k + J_{IC}(\bar{N}_k) \sum_i \frac{|w_{k0i}|^2}{(|w_{k0i}| + \eta)^2}, \end{aligned} \quad (131)$$

$$\xi_{kn} = y_n \mathbf{w}_k^\dagger \mathbf{x}_n, \quad (132)$$

$$W_k = W_k(\mathbf{w}_{k0}) = 2 \sum_n r_{nk} \lambda(\xi_{kn}) \mathbf{x}_n \mathbf{x}_n^\dagger, \quad (133)$$

$$\mathbf{s}_k = \frac{1}{2} \sum_n r_{nk} y_n \mathbf{x}_n, \quad (134)$$

$$(\mathbf{a}_k)_i = J_{IC}(\bar{N}_k) \frac{\eta}{(|w_{k0i}| + \eta)^2}, \quad (135)$$

$$C_k = C_k(\mathbf{w}_{k0}) = \sum_n r_{nk} \left(\frac{1}{2} \xi_{kn} - \log \sigma(\xi_{kn}) - \lambda(\xi_{kn}) \xi_{kn}^2 \right). \quad (136)$$

As discussed in “Proposed Method: SICM Algorithm”, we can approximately minimize F_k and estimate the parameters of the k th mixture component by minimizing the upper bound $\mathcal{F}_k$ in Eq. (131) with respect to $\mathbf{w}_k$ and $\mathbf{w}_{k0}$. We can minimize $\mathcal{F}_k$ with respect to $\mathbf{w}_k$ using the subgradient method explained in Appendix C.

From the discussion presented above, the parameter estimation algorithm of the k th logistic regression component is summarized as Algorithm 4. The SICM algorithm for solving the HRD problem in the case of a logistic mixture is constructed by replacing Step 5 of Algorithm 2 with Algorithm 4 ($k = 1, 2, \dots, K$).

Algorithm 4 SICM algorithm for the k -th component of a logistic mixture.

1. Setting. Choose AIC or BIC as the information criterion to be minimized. Input the user-defined parameters: η (positive constant to approximate an L_0 norm) and ϵ (threshold for convergence condition used in Step 5 below).
2. Initialization. Input the training observations (Y, X) and the expected assignment probability $\{r_{nk}\}$. Represent an update step as s and set $s = 0$. Conduct maximum likelihood estimation and define its estimator as $\mathbf{w}_k^{(s=0)} = \mathbf{w}_{k0}^{(s=0)}$. Estimate ξ_k , W_k , $\mathbf{s}_k$, $\mathbf{a}_k$, and C_k . Estimate the objective function: $\mathcal{F}_k^{(s=0)} = \mathcal{F}_k(\mathbf{w}_k^{(s=0)}, \mathbf{w}_{k0}^{(s=0)})$.
3. Update of the parameters. Update $\mathbf{w}_k$ by solving the following minimization problem:

$$\mathbf{w}_k^{(s+1)} = \underset{\mathbf{w}_k}{\operatorname{argmin}} \frac{1}{2} \mathbf{w}_k^\dagger W_k \mathbf{w}_k - \mathbf{s}_k^\dagger \mathbf{w}_k + \|\mathbf{a}_k \circ \mathbf{w}_k\|_1. \quad (137)$$

This problem is solvable using the subgradient method explained in Appendix C. Update $\mathbf{w}_{k0}$ as $\mathbf{w}_{k0}^{(s+1)} = \mathbf{w}_k^{(s+1)}$.

4. Update other than the parameters. Update ξ_k , W_k , $\mathbf{a}_k$, and C_k ($\mathbf{s}_k$ is not updated because it is independent of $\mathbf{w}_{k0}$). Estimate the objective function: $\mathcal{F}_k^{(s+1)} = \mathcal{F}_k(\mathbf{w}_k^{(s+1)}, \mathbf{w}_{k0}^{(s+1)})$.
5. Convergence determination. If the following convergence condition

$$\mathcal{F}_k^{(s)} - \mathcal{F}_k^{(s+1)} < \epsilon \quad (138)$$

is satisfied, then go to Step 6. Otherwise, set $s = s + 1$ and go to Step 3.

6. Output the results and stop the algorithm.

References

1. Corduneanu A, Bishop CM. Variational Bayesian model selection for mixture distributions. In: Proceedings 8th international conference on artificial intelligence and statistics, pp 27–34 (2001)
2. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em algorithm. *J R Stat Soc Ser B*. 1977;39(1):1–38.
3. Akaike H. Information theory and an extension of the maximum likelihood principle. In: 2nd international symposium on information theory. Akademia Kiado, Budapest. 1973 p. 267–281.
4. Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6:461–4.
5. Attias H. Inferring parameters and structure of latent variable models by variational Bayes. In: Proceedings of the fifteenth conference on uncertainty in artificial intelligence, UAI'99. 1999. p. 21–30.
6. Teh YW, Newman D, Welling M. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In: Advances in neural information processing systems, vol 19. MIT Press; 2007. p. 1353–1360
7. Fujimaki R, Morinaga S. Factorized asymptotic Bayesian inference for mixture models. In: AISTATS. 2012.
8. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. Berlin: Springer; 2001.
9. Zhang T. Adaptive forward-backward greedy algorithm for sparse learning with linear models. In: Proceedings of the 21st international conference on neural information processing systems. 2008. p. 1921–28.
10. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B (Methodological)*. 1996;39(1):267–88.
11. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001;96(456):1348–60.
12. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat*. 2004;32(2):407–99.
13. Eto R, Fujimaki R, Morinaga S, Tamano H. Fully-automatic Bayesian piecewise sparse linear models. In: AISTATS. 2014.
14. Nguyen TNA, Bouserdoum A, Phung SL. Scalable hierarchical mixture of gaussian processes for pattern classification. In: IEEE international conference on acoustics, speech and signal processing. 2018. p. 2466–70.
15. Fan K, Li C, Heller K. A unifying variational inference framework for hierarchical graph-coupled hmm with an application to influenza infection. In: 13th AAAI conference on artificial intelligence. 2016.
16. Laufer Y, Gannot S. A Bayesian hierarchical model for speech dereverberation. In: IEEE international conference on the science of electrical engineering. 2018. p. 1–5.
17. Azzimonti L, Corani G, Zaffalon M. Hierarchical multinomial-Dirichlet model for the estimation of conditional probability tables. In: IEEE international conference on data mining (ICDM). 2017. p. 733–44.
18. Hunter DR, Lange K. A tutorial on mm algorithms. *Am Stat*. 2004;58(1):30–7.
19. Bache K, Lichman M. UCI machine learning repository. 2013. <http://archive.ics.uci.edu/ml>. Accessed 27 Feb 2020.
20. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. 2006;101(476):1418–29.
21. scikit-learn. <https://scikit-learn.org/stable/>. Accessed 27 Feb 2020.
22. Drugowitsch J. Variational Bayesian inference for linear and logistic regression. 2013. [arXiv:1310.5438](https://arxiv.org/abs/1310.5438).
23. Bishop CM. Pattern recognition and machine learning. Berlin: Springer; 2006.
24. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. 2008;9(3):432–41.
25. Jaakkola TS, Jordan MI. Bayesian parameter estimation via variational methods. *Stat Comput*. 2000;10(1):25–37.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.