Noname manuscript No.
(will be inserted by the editor)

Hybrid Adaptive Index Model for Binary Response Data

Ke Wan[1] · **Kensuke Tanioka**[3]
· **Hiroyuki Minami**[1] · **Masahiro Mizuta**[1]
· **Toshio Shimokawa**[2]

Received: date / Accepted: date

Abstract We often meet the case in data analysis that the explanatory variables can be occasionally divided into two groups. One group comprises the variables that researchers consider controllable, and the other group comprises those they do not. We call them controllable and uncontrollable variables, respectively. In the study, we deal with binary response data and aim to estimate the relationship between the binary response and controllable variables. Logistic regression model is typically used in binary response data. In addition to that, AIM (Adaptive Index Model; Tian and Tibshirani, 2010) can also be used in binary response data. Contrast with logistic regression model, AIM can explain the result easier using binary rules but the prediction accuracy of AIM is shown worse than that of logistic regression model. Considering the interpretability and accuracy, it is better to apply AIM to controllable variables and adjust the effect of uncontrollable variables using logistic regression model. Therefore, we propose the method combining AIM and logistic regression model, called hybrid adaptive index model(HAIM), to give best solution.

Keywords Production rule · Logistic regression model · controllable explanatory variables · uncontrollable explanatory variables

Ke Wan
E-mail: wane19911017@gmail.com

Kensuke Tanioka
E-mail: ktaniok@mail.doshisha.ac.jp

Hiroyuki Minami
E-mail: min@iic.hokudai.ac.jp

Masahiro Mizuta
E-mail: mizuta@iic.hokudai.ac.jp

Toshio Shimokawa
E-mail: toshibow2000@gmail.com

1. Graduate School of Information Science and Technology, Hokkaido University. ·
2. Department of Medical Data Science, Graduate School of Wakayama Medical University. ·
3. Faculty of Life and Medical Sciences, Doshisha University

1 Introduction

In this paper we focus on the situation that explanatory variables in binary response data can be divided into controllable variables and uncontrollable variables. Controllable variables are the variables which can be changed or controlled, such as salt intake and alcohol intake. Uncontrollable variables are the variables which are not controllable in general, such as sex and age. For example, in medical field, the best way to prevent cerebral infraction is to manage each risk factor and keep in a good condition. However, not all the risk factors are controllable, e.g. age is a known risk factor but is uncontrollable. The management of the controllable variables such as salt intake and alcohol intake seems to be much more important. Therefore, it is significant to derive the interpretable attributes of subjects based on controllable variables and detect the relationship between the response and these attributes of subjects.

In order to detect the relationship between the attributes of subjects and response, the classification methods or subgroup identification methods can be used. For example, the tree-based methods such as CART (Classification and Regression Trees; Breiman et al., 1984), IT (Interaction Trees; Su et al., 2009) and SIDES (Subgroup Identification Based on Difference Effect Search; Lipkovich et al., 2011). Additionally, Li et al. (2018) proposed the method to detect multiple-threshold and Wang et al. (2019) threshold based on the combination of multivariate predictors which can derive more meaningful partition. Furthermore, in this paper we concentrate to derive the interpretable attributes of subjects based on controllable variables. Thus, the explanatory variables need to be divided at first. However, the tree-based methods get the results based on all of the explanatory variables and cannot avoid the interaction between the controllable variables and uncontrollable variables, so that these methods are difficult to get the results based only on controllable variables. Although the methods proposed by Li et al. (2018) and Wang et al. (2019) can derive the meaningful attributes of the subjects from multiple-thresholds and the threshold based on the combination of multivariate predictors, the interpretation of results could be difficult. Beside these methods, Tian and Tibshirani (2010) proposed AIM (Adaptive Index Model) which can get a sequence of simple indices (e.g. "Blood pressure $>$ 120 mm Hg") based on explanatory variables. Because of each index is constructed independently, therefore we can easily get the results based only on controllable variables. However, the Tian and Tibshirani (2010) also shows that the performance of AIM tends to be bad, when the data tends to be linear.

There are several studies discussed AIM, Chen et al. (2015) compare the performance of the PRIM-based method with AIM, Huang et al. (2017) modify the AIM for subgroup identification and Shimokawa et al. (2013) use the residual deviance to estimate each parameter simultaneously, those methods are not considering to obtain the result based only on controllable variables considering the effect of uncontrollable variables. Therefore, we propose a method to deal with such situation. Here, we combine AIM with logistic regression model called hybrid adaptive index model (HAIM). In HAIM, AIM is used to construct binary rules from controllable variables and logistic regression model is applied to adjust the effect of uncontrollable variables. We expect that the prediction accuracy of HAIM tends to be good through the logistic regression model of uncontrollable variables. In this way, we can get the re-

sult based only on controllable variables, and interpret the feature corresponding to controllable variables with considering the effect of uncontrollable variables.

The remainder of this paper is organized as follows. In Section 2, we introduce AIM and its algorithm. In Section 3, we propose HAIM and its algorithm in detail. In Section 4, we conduct numerical simulations to compare the accuracy of HAIM, AIM, and the CART. We compare HAIM with AIM using a real example and interpret the features of the real example using HAIM in Section 5. In Section 6, we summarize the paper.

2 AIM

In this section, AIM is introduced, which is proposed by Tian and Tibshirani (2010). AIM constructs a set of binary rules adaptively such as "$x_j > c_j$" or "$x_j < c_j$" based on explanatory variables of data. For each observation, a score called index predictor, can be obtained from the sum of binary scores corresponding to the set of binary rules. The higher index predictor can be easily interpreted as the higher risk of event. The advantage of AIM is easy to interpret and can capture the situation where the effect to response relate to multiple variables.

AIM is shown in detail as follow. Let $\boldsymbol{x}$ a vector consist of p explanatory variables, $\boldsymbol{x} = (x_1, \cdots, x_p)$. Here, the sample is given of N observations $\{y_i, x_{ij}\}_1^N (j = 1, 2, \cdots, p)$. The model consists of K indicator functions defined as follows:

$$\log(\pi_i/1-\pi_i) = \beta_0 + \beta_1 \sum_{j \in C_K} I(s_j x_{ij} < s_j c_j) \tag{1}$$

where β_0 and β_1 are the intercept and the coefficient, respectively, $\pi_i = \Pr\{y_i = 1 | \boldsymbol{x}_i\}$ is the probability of $y_i = 1$, given $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})$ where $y_i \in \{0, 1\}$ is a binary response. $I(s_j x_{ij} < s_j c_j)$ is the indicator function for which $I(s_j x_{ij} < s_j c_j)$ is one if $s_j x_{ij} < s_j c_j$ is true, and $I(s_j x_{ij} < s_j c_j)$ is zero, otherwise. $s_j \in \{-1, 1\}$ is the sign of jth explanatory variable, C_K is the set of selected indices of variables $C_K \subset \{1, 2, \cdots, p\} (|C_K| = K)$ and $c_j (j \in C_k)$ is the cutoff point of jth variable. The explanatory variable $\boldsymbol{x}_j$ and the corresponding cutoff point c_j is selected in a forward stepwise manner to maximized the score test statistics under the null hypothesis, $\beta_1 = 0$ in Eq. (1). For the details of the Algorithm, see Tian and Tibshirani (2010).

3 HAIM

In this section, we propose hybrid adaptive index model (HAIM) for binary response. AIM can use several binary rules to explain the relationship between the binary response and the explanatory variables. In addition to that, we can get index predictors according to the sum of binary scores corresponding to the set of binary rules and estimate the risk of event. Contrast to logistic regression model, AIM make the result easy to interpret. However, the prediction accuracy of AIM shows worse than logistic regression (Tian and Tibshirani, 2010). As we have mentioned, a situation is considered that explanatory variables in data can be divided into controllable variables

and uncontrollable variables. Here, we only concentrate on controllable variables and construct a set of binary rules based entirely on controllable variables. The effect of uncontrollable variables should also be considered. We use logistic regression to adjust the effect of uncontrollable variables and expect that the prediction accuracy of HAIM tends to be better. Consequently, we combine AIM with logistic regression model.

Model and Algorithm

We designate $\boldsymbol{x}$ a vector consist of p explanatory variables, $\boldsymbol{x} = (\boldsymbol{x}^{\mathrm{c}}, \boldsymbol{x}^{\mathrm{uc}}) = (x_1^{\mathrm{c}}, \cdots, x_m^{\mathrm{c}}, x_1^{\mathrm{uc}}, \cdots, x_n^{\mathrm{uc}})(m+n=p)$ where $\boldsymbol{x}^{\mathrm{c}} = (x_1^{\mathrm{c}}, \cdots, x_m^{\mathrm{c}})$ are controllable variables and $\boldsymbol{x}^{\mathrm{uc}} = (x_1^{\mathrm{uc}}, \cdots, x_n^{\mathrm{uc}})$ are uncontrollable variables. Here, the sample is given of N observations $\{y_i, \boldsymbol{x}_i\}_1^N$ where $\boldsymbol{x}_i = (x_{ij}^{\mathrm{c}}, x_{il}^{\mathrm{uc}}) = (x_{i1}^{\mathrm{c}}, \cdots, x_{im}^{\mathrm{c}}, x_{i1}^{\mathrm{uc}}, \cdots, x_{in}^{\mathrm{uc}})(m+n=p)$. We would like to construct binary rules based solely on controllable variables $\boldsymbol{x}^{\mathrm{c}} = (x_1^{\mathrm{c}}, \cdots, x_m^{\mathrm{c}})$, considering the effects of uncontrollable variables $\boldsymbol{x}^{\mathrm{uc}} = (x_1^{\mathrm{uc}}, \cdots, x_n^{\mathrm{uc}})$. We define HAIM for binary response as follows:

$$\log(\pi_i/(1-\pi_i)) = \beta_0 + \beta_1 \sum_{j \in C_K} I(s_j x_{ij}^{\mathrm{c}} < s_j c_j) + \sum_{l=1}^{n} \alpha_l x_{il}^{\mathrm{uc}} \tag{2}$$

where β_0, β_1 and $\alpha_l (l = 1, \cdots, n)$ are the intercept, coefficients for the controllable variables, and coefficients for the uncontrollable variables, respectively. Here, $\pi_i = \Pr\{y_i = 1 | \boldsymbol{x}_i^{\mathrm{c}}, \boldsymbol{x}_i^{\mathrm{uc}}\}$ is the probability of $y_i = 1$, given $\boldsymbol{x}_i^{\mathrm{c}} = (x_{i1}^{\mathrm{c}}, \cdots, x_{im}^{\mathrm{c}})$ and $\boldsymbol{x}_i^{\mathrm{uc}} = (x_{i1}^{\mathrm{uc}}, \cdots, x_{in}^{\mathrm{uc}})$ where $y_i \in \{0, 1\}$ is a binary response. C_K is the set of selected indices of controllable variables $C_K \subset \{1, 2, \cdots, m\}(|C_K| = K)$, and c_j is the cutoff point of variable $j (j \in C_K)$. In this model, we adjust the effects of the uncontrollable variables $\boldsymbol{x}_i^{\mathrm{uc}}$ using logistic regression model and apply AIM only to controllable variables $\boldsymbol{x}_i^{\mathrm{c}}$. In HAIM, we can estimate the optimal c_j maximized the score test statistics of test under the null hypothesis, $\beta_1 = 0$ in Eq.(2). HAIM algorithm which estimate the optimized set of controllable variables $C_K \subset \{1, 2, \cdots, m\}(|C_K| = K)$ and corresponding cutoff point c_j is shown as follow.

Algorithm 1 HAIM algorithm

1: Set K as the number of binary rules, $V = \{1, 2, \cdots, m\}$ as the set of the indices of controllable variables and $C_K \leftarrow \{\phi\}$.
2: Estimate β_0, $\alpha_l (l = 1, \cdots, n)$ with the response variable y and uncontrollable variables $\boldsymbol{x}^{\mathrm{uc}}$.
3: Set new response variable y^*, where $y_i^* = \log(\pi_i/(1-\pi_i)) - (\beta_0 + \sum_{l=1}^{n} \alpha_l x_{il}^{\mathrm{uc}})$.
4: **while** $|C_K| \leq K$ **do**
5: For all $k \in V$, determine c_k, s_k such that the corresponding score test statistic of test $\beta = 0$ is maximized, given y^*, x^{c} and x^{uc}
6: Let k^* be the controllable variables which maximizing the score test statistic among $k \in V$ and $C_K \leftarrow C_K \bigcup \{k^*\}$ and $V \leftarrow V \backslash \{k^*\}$
7: **end while**

4 Numerical Studies

In this section, several simulations are performed to evaluate the performances of HAIM, AIM and CART by prediction accuracy under the various situations.

4.1 Simulations for the prediction accuracy

Here, we show the results of numerical simulations for evaluating the prediction accuracy. We focus on the data $\{y, \boldsymbol{x}^{\mathrm{c}}, \boldsymbol{x}^{\mathrm{uc}}\}$ with binary response $y \in \{0,1\}$ and the explanatory variables can be divided into controllable variables $\boldsymbol{x}^{\mathrm{c}}$ and uncontrollable variables $\boldsymbol{x}^{\mathrm{uc}}$. There are several datasets generated for simulation study and the simulation settings will be explained later.

Explanatory variables settings

The explanatory variables $\boldsymbol{x} = (\boldsymbol{x}^{\mathrm{c}}, \boldsymbol{x}^{\mathrm{uc}})$ consist of the controllable variables $\boldsymbol{x}^{\mathrm{c}} = (x_1^{\mathrm{c}}, x_2^{\mathrm{c}}, x_3^{\mathrm{c}})$ and uncontrollable variables $\boldsymbol{x}^{\mathrm{uc}} = (x_1^{\mathrm{uc}}, x_2^{\mathrm{uc}}, x_3^{\mathrm{uc}})$. Each explanatory variable is distributed from standard normal distribution that $x_1^{\mathrm{c}}, x_2^{\mathrm{c}}, x_3^{\mathrm{c}}, x_1^{\mathrm{uc}}, x_2^{\mathrm{uc}}, x_3^{\mathrm{uc}} \sim N(0,1)$. Considering the correlation between the explanatory variables, the explanatory variables are generated in four cases:

Case 1: Each explanatory variable is generated independently.
Case 2: A correlation between x_1^{c} and x_1^{uc} with $r = 0.6$
Case 3: A correlation between x_1^{c} and x_2^{c} with $r = 0.6$
Case 4: A correlation between x_1^{uc} and x_2^{uc} with $r = 0.6$

Response variables settings

The response variable $y \in \{0,1\}$ is dichotomous. In order to confirm the prediction accuracy of proposed method in different situation, we set five models to generate the response variables and set $e \sim (0,1)$ as error term for each model. Model 1 (Eq. 3), Model 2 (Eq. 4) and Model 3 (Eq. 5) are step functions. In model 1, the situation is set in favor to AIM and split of each explanatory variable is only one. However, Model 2 set two splits and Model 3 set four splits for each explanatory variable. Model 4 (Eq. 6) is linear model. Model 5 (Eq. 7) is set in favor to HAIM and combine step function and linear model. In model 5, the step function which is similar to the Model 1 is applied to controllable variables and the linear model is applied to uncontrollable variables.

Model 1:

$$f(x) = 1 - 0.5 \sum_{j=1}^{3} I(x_j^{\mathrm{c}} > 0) - 0.5 \sum_{l=1}^{3} I(x_l^{\mathrm{uc}} > 0) + 0.5e \tag{3}$$

Model 2:

$$\begin{aligned} f(x) = \ 5 \ & -0.5 \sum_{j=1}^{3} [I(x_j^{\mathrm{c}} > -1) + I(x_j^{\mathrm{c}} > 1)] \\ & - \ 0.5 \sum_{l=1}^{3} [I(x_l^{\mathrm{uc}} > -1) + I(x_l^{\mathrm{uc}} > 1)] + 0.5e \end{aligned} \tag{4}$$

Model 3:

$$f(x) = 7 - 0.5\sum_{j=1}^{3}[I(x_j^{c} > -1) + I(x_j^{c} > -0.5) + I(x_j^{c} > 0.5) + I(x_j^{c} > 1)] - 0.5\sum_{l=1}^{3}[I(x_l^{uc} > -1) + I(x_l^{uc} > -0.5) + I(x_l^{uc} > 0.5) + I(x_l^{uc} > 1)] + 0.5e \quad (5)$$

Model 4:

$$f(x) = -0.5\sum_{j=1}^{3} x_j^{c} - 0.5\sum_{l=1}^{3} x_l^{uc} + 0.5e \quad (6)$$

Model 5:

$$f(x) = 1 - 0.5\sum_{j=1}^{3} I(x_j^{c} > 0) - 0.5\sum_{l=1}^{3} x_l^{uc} + 0.5e \quad (7)$$

According to these generation model the $f(x)$ is calculated and then the binary response variable y is generated by

$$y = \begin{cases} 1, & 1/[1+exp(-f(x))] > 0.5 \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

Size of training and test data

In the simulation, the size of trainings data is set as 100, 200, 300, 400 and 500. The size of test data is set as 200. In addition, the summary of the simulation design is shown in Table 1.

Table 1 The summary of simulation settings

Explanatory variables	Case 1, Case 2, Case 3 and Case 4
Response variables	Model 1, Model 2, Model 3, Model 4 and Model 5
Training data size	100, 200, 300, 400 and 500
Test data size	200

Then, three methods AIM, HAIM and CART are applied to each dataset. In HAIM we have to divide explanatory variables into controllable variables and uncontrollable variables at first. Therefore, we set x_1^{uc}, x_2^{uc} and x_3^{uc} as observed uncontrollable variables and the other explanatory variables as controllable variables. Then, the performance of HAIM will be compared with that of AIM and CART which are evaluated using prediction accuracy. The prediction accuracy is calculated by the correct classification rate between true binary responses y and estimated binary responses $\hat{y}$ of test data and can be describe as follow.

$$\text{correct classification rate} = \frac{\sum_1^{200} I(\hat{y}_i = y_i)}{200} \quad (9)$$

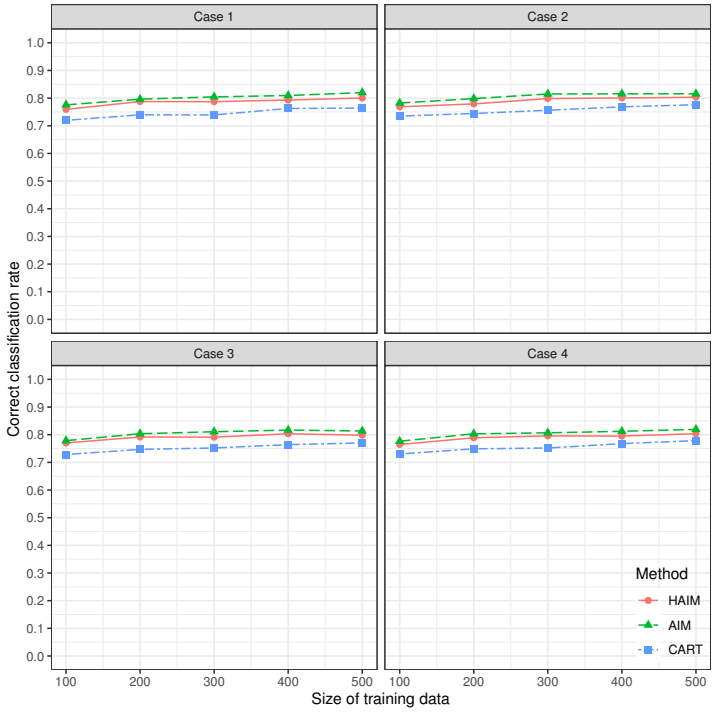


Fig. 1 The results of simulation data generated by model 1

Here, the $\hat{y}$ is calculated in the same way as Eq. 8, where the $f(x)$ is estimated HAIM. The test data is constructed in the same way as each of the corresponding training data.

4.2 The results of simulation studies for the prediction accuracy

Simulation data generated by model 1

The results of simulation data generated by model 1 is shown as Fig. 1. As mentioned above, this simulation data is favor to AIM and AIM shows higher correct classification rates than those of HAIM and CART. However, the performances of HAIM are not so worse than AIM and both the correct classification rates of AIM and HAIM are nearly 0.8. Case 1 to Case 4 show the similar results. Therefore, the cases of the correlation between x_1^{c} and x_1^{uc} with $r = 0.6$, the correlation between x_1^{c} and x_2^{c} with $r = 0.6$ and the correlation between x_1^{uc} and x_2^{uc} with $r = 0.6$ shows the similar correct classification rates as the case of each explanatory variable is generated independently. Additionally, for each method the correct classification rates tend to be higher with increasing size of the training data.

Simulation data generated by model 2, model 3 and model 4

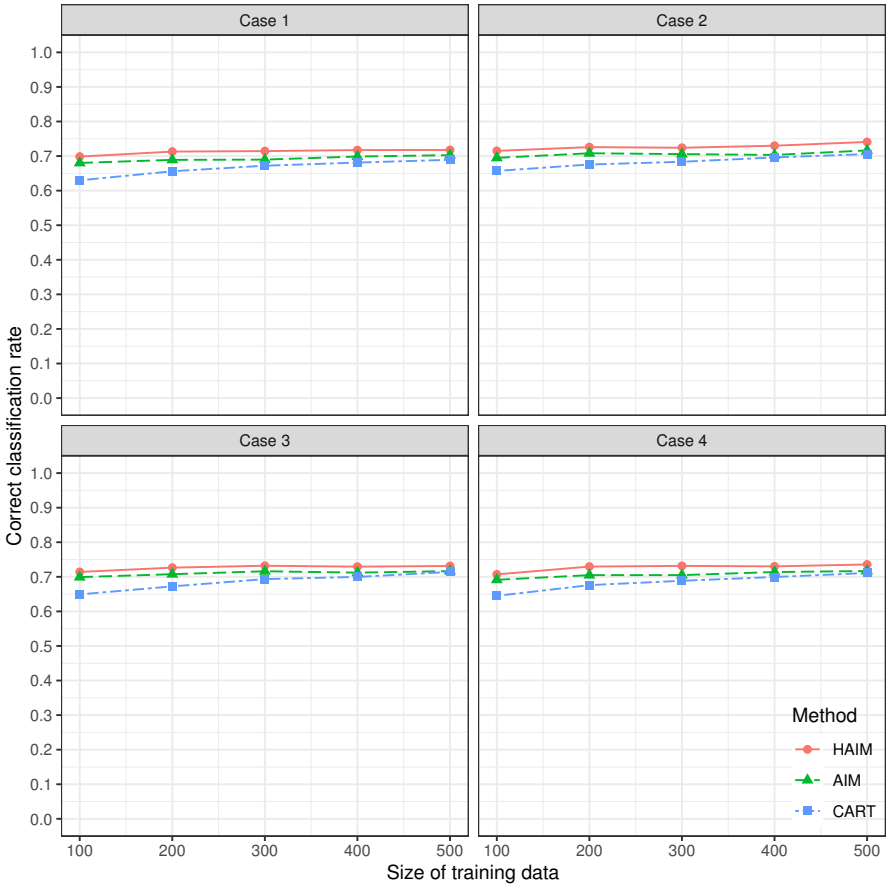


Fig. 2 The results of simulation data generated by model 2

The results of simulation data generated by model 2, model 3 and model 4 is shown in Fig. 2, Fig. 3 and Fig. 4. The Model 2 and Model 3 is step function. However, unlike the Model 1, the Model 2 and Model 3 sets multi-threshold for each explanatory variable. Therefore, the Model 2 and Mode 3 tends to be more linear than the Model 1. The Model 4 is linear model. The correct classification rates of HAIM tend to be higher than those of AIM and CART in these simulation data. Especially, we the data is generated by Model 3 and Model 4, the correct classification rates of HAIM are high and almost over 0.8. Furthermore, the correct classification rates of Case 2, Case 3 and Case 4 tends to be higher than those of Case 1. Hence, the cases of the correlation between x_1^c and x_1^{uc} with $r = 0.6$, the correlation between x_1^c and x_1^c with $r = 0.6$ and the correlation between x_1^{uc} and x_1^{uc} with $r = 0.6$ can improve the correct classification rates of each method. Same as above, for each methods the size of training data is large, the correct classification rate tends to be higher.

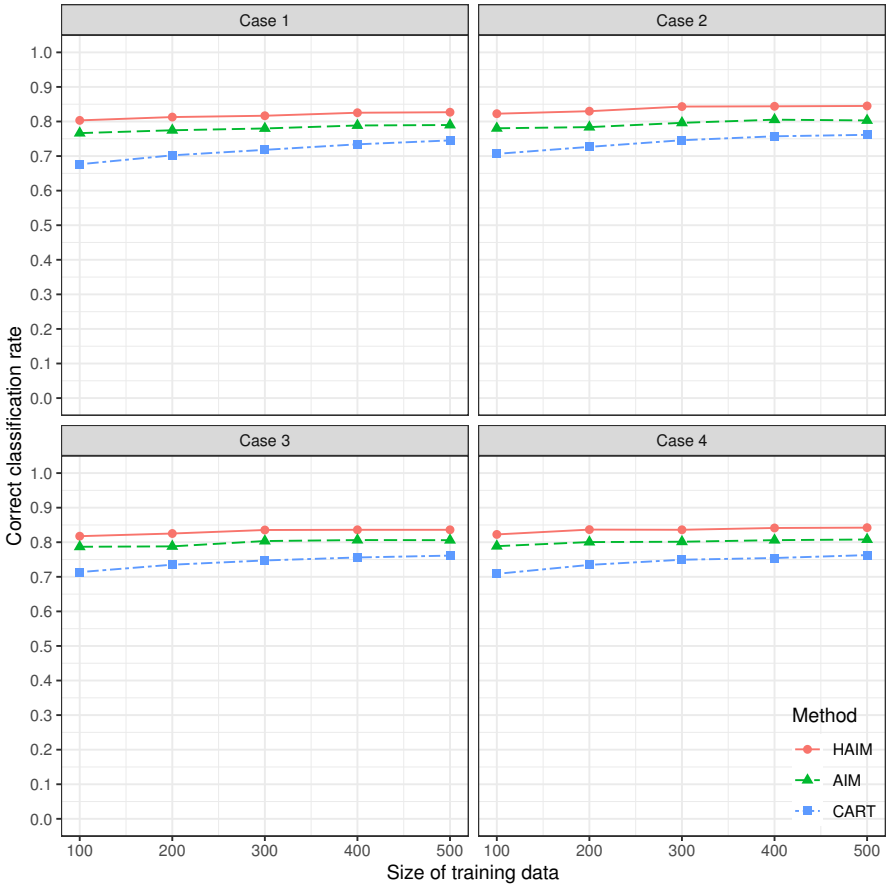


Fig. 3 The results of simulation data generated by model 3

Simulation data generated by model 5

The results of simulation data generated by model 5 is shown in Fig. 5. As explained above, this simulation data is favor to HAIM and HAIM shows much higher correct classification rates than those of AIM and CART. Especially, when the size of training data is over 300, the correct classification rates almost achieve 0.85 and better than the data generated by other models. Moreover, the correct classification rates of Case 1 and Case 2 are similar, but those of Case 2 and Case 4 are higher than those of Case 1 and Case 2. It shows that when the true model is model 5, the correlation between x_1^c and x_1^c with $r = 0.6$ lead the same results as the each explanatory variable is generated independently. Additionally the correlation between x_1^c and x_1^c with $r = 0.6$ and the correlation between x_1^{uc} and x_1^{uc} with $r = 0.6$ improve the correct classification rates when the generation model is model 5. Also, the large size of training data can improve the correct classification rates for each method.

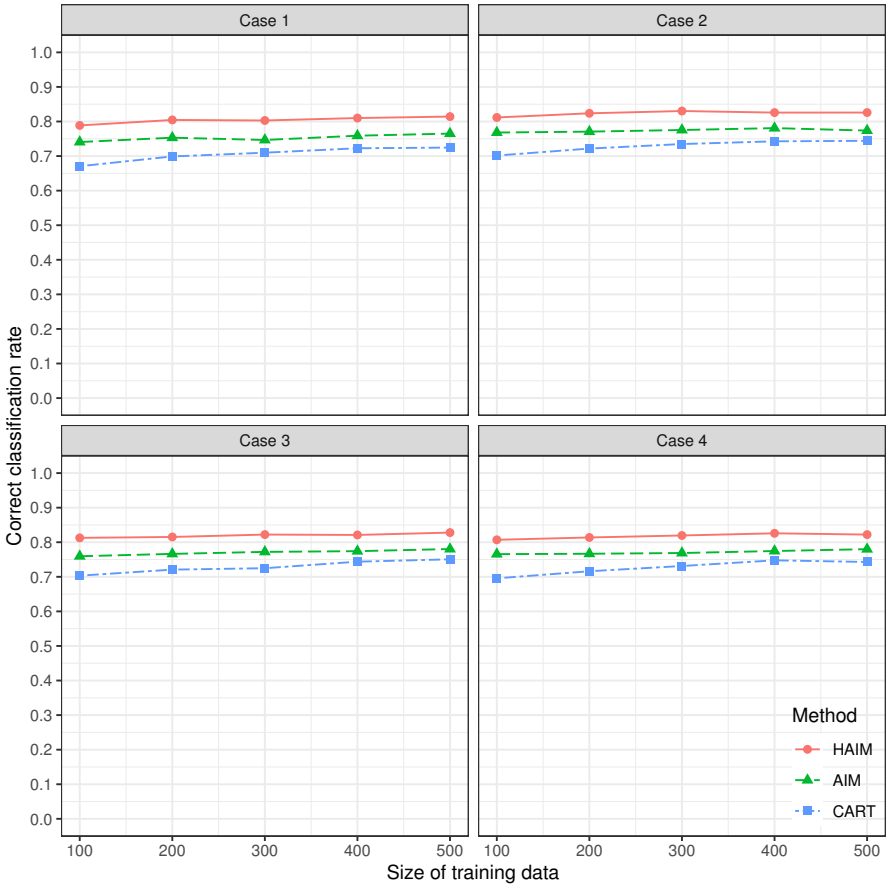


Fig. 4 The results of simulation data generated by model 4

Summary

According to these simulations, we can see the correct classification rates of AIM are little higher than those of HAIM when the simulation data are generated by model 1. However, when the simulation data are generated by model 2, model 3, model 4 and model 5 the correct classification rates of HAIM are higher than those of AIM. The model 1 is favor to AIM and model 5 is favor to HAIM. Therefore, the correct classification rates of AIM are highest in model 1 and those of HAIM are highest in model 5. The model 2 and model 3 are step functions and the model 4 is linear models, it shows that the correct classification rates of HAIM tend be higher when the generation model tend to be linear, but those of AIM tend to be lower.

Concentrate on the correlation within the explanatory variables. In model 1, the correct classification rates of HAIM are not affected by the correlation within the explanatory variables. In model 2, model 3 and model 4 the correlation between x_1^c and x_1^{uc} with $r = 0.6$, the correlation between x_1^c and x_1^c with $r = 0.6$ and the correlation

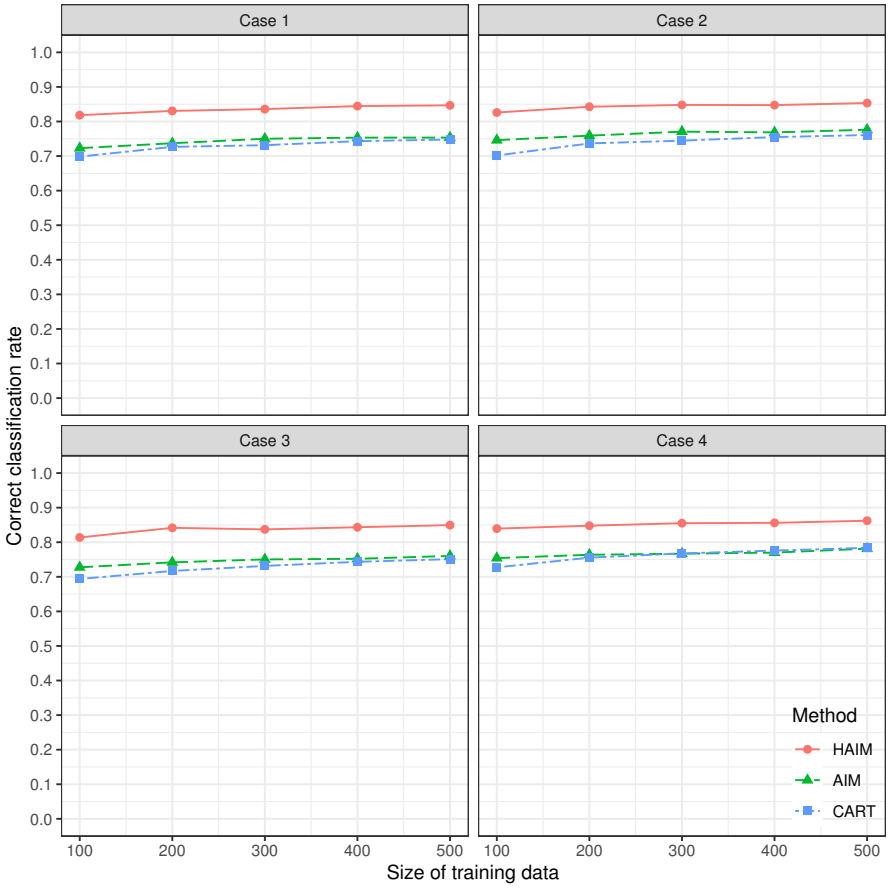


Fig. 5 The results of simulation data generated by model 5

between x_1^{uc} and x_1^{uc} with $r = 0.6$ improve the correct classification rates of HAIM. In model 5, only the correlation between x_1^{c} and x_1^{uc} with $r = 0.6$ and the correlation between x_1^{c} and x_1^{c} with $r = 0.6$ effect the correct classification rates of HAIM. We can see that in model 1 each variable is assigned to the one index functions and in model 5 each controllable variable is assigned to one index function.

Summarize the results of the simulation, we can see the HAIM have two features:

- the correct classification rate of HAIM is higher than AIM when the structure of data tends to be linear.
- the correlation within the explanatory variables can improve the correct classification rate of HAIM when the structure of data tends to be linear.

5 Real example

In this section, through a dataset of real example, we show the interpretability of the result of HAIM, AIM and CART. The result of AIM and HAIM will be shown and the result of CART will be provided to compare the result of AIM and HAIM from the perspective of the interpretation.

The dataset is a case study using a survey regarding tourist satisfaction in several ancient towns in Chengdu, China provided by Liu et al. (2017). The aim of the survey is to interpret the influence factors of tourist satisfaction. The sample size is 545, and the tourist satisfaction of the ancient town evaluated through "satisfaction" and "unsatisfaction" as the binary response variable. Table 2 contains 26 influence variables that are considered as explanatory variables related to the binary response. These variables are evaluated by five-segment rating method.

We would like to get a set of indices which can help us to distinguish whether a tourist is satisfied with the ancient town. According to that we can easily calculate the number of satisfied tourists and know the influence variables mainly contribute to the satisfaction which can give us useful information about tourist satisfaction of ancient town. In addition to that, it is necessary to consider whether this information can be used to improve tourist satisfaction, because it is impossible to improve tourist satisfaction even if we could interpret the features of variables that do not lend themselves to intervention. For example, both the landscape and cleanliness of ancient town can influence the satisfaction of tourists. The cleanliness of ancient town can be easily improved by cleaning. However, the landscape of ancient town is always strictly protected by policy and impossible to be changed by intervention. Therefore, we would like to determine indices based solely on controllable variables.

We divided the explanatory variables into controllable variables and uncontrollable variables. For the classification of these variables, see Table 2. In this real example, controllable variables are considered as capable of being influenced by intervention, such as the cleanliness of the tourist spot, the kind of amusement provided, and the expression of traditional life. Uncontrollable variables are considered as incapable of being influenced by intervention.

We apply HAIM and AIM to the data and estimate the optimal number of binary rules K using tenfold cross-validation. Figure 6 shows the cross-validated score test statistics for each K of HAIM and AIM where cross-validated score test statistics is calculated through the mean of test statistics for each K. In HAIM $K = 2$ and in AIM $K = 3$ is selected since the corresponding cross-validated score test statistics is highest. Therefore, we apply HAIM with $K = 2$ and AIM with $K = 3$.

The result of the application is shown in Table 3. We can see that both the indices of HAIM and AIM consist of "Q1. Enjoyment of Local Life", "Q2. Verbal Introduction to Traditional Life". It indicates that the tourists of ancient towns would like to enjoy the traditional life. The"Q1. Enjoyment of Local Life" is evaluated quite strictly and only the tourists who select 4 or 5 are satisfied. It shows that tourists require the high quality of "Q2. enjoyment of Local Life ". However, the cut off value of "Q2. Verbal Introduction to Traditional Life" is 1 and can be easily satisfied. Compare the difference of HAIM and AIM, we can see the indices of HAIM only select controllable variables while the indices of AIM select "Q7. Historical Buildings>3" the

Table 2 The question of influence factors of satisfaction in Liu et al. 2017

Controllable variables
Q1. Enjoyment of Local Life
Q2. Verbal Introduction to Traditional Life
Q3. Experience of Local Traditional Life
Q4. Enjoyment of Tea Houses
Q5. Communication with Local Residents
Q6. Enjoyment of the Sichuan Opera
Q10. Green Vegetative Landscapes
Q12. Education on Local History and Culture
Q13. Enjoyment of Traditional Delicacies
Q14. Enjoyment of Non-Traditional Foods
Q15. Availability of Traditional Objects for Sale
Q16. Availability of Non-Traditional Items for Sale
Q17. Leisure and Rest
Q18. Enjoyment of Mahjong
Q19. Reunion of Family and Friends
Q20. Encounter
Q21. Street Cleanliness
Q22. Bathroom Cleanliness
Q23. Street Comfort
Q24. Enjoyment of Rest Stops
Q25. Commentary on Streets
Q26. Enjoyment of Night Life
Uncontrollable variables
Q7. Historical Buildings
Q8. Traditional Street Landscapes (Overall Atmosphere)
Q9. Waterside Landscapes
Q11. Historical Anecdotes
Response
Q. Tourists satisfaction

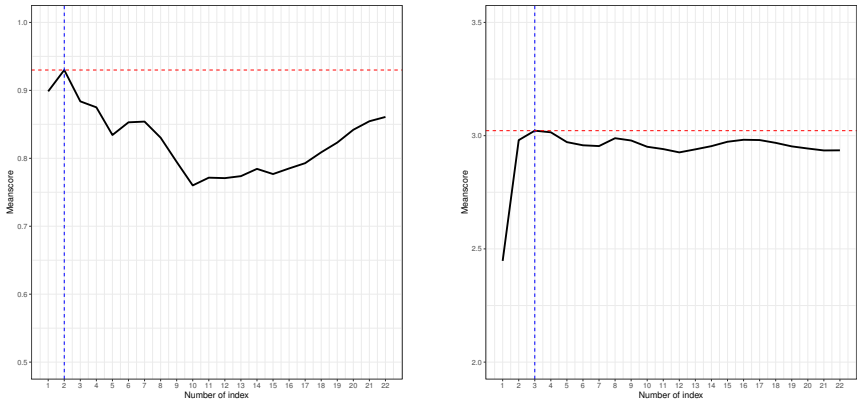
Table 3 Comparison between HAIM and AIM

Method	**Question**	**Threshold**
HAIM	Q1. Enjoyment of Local Life	>3
	Q2. Verbal Introduction to Traditional Life	>1
AIM	Q1. Enjoyment of Local Life	>3
	Q2. Verbal Introduction to Traditional Life	>1
	Q7. Historical Buildings	>3

landscape of ancient town which are impossible to intervene. According to the above results, the traditional life and landscape is important for the tourists. Both "Enjoyment of Local Life" and "Historical Buildings" are strictly evaluated. Although the "Verbal Introduction to Traditional Life" is the influence factors of tourist satisfaction, tourists are not care about its quality and can be satisfied easily. In addition, we calculate the percentage of subjects who are satisfied with the binary rules of selected controllable variables. The percentage of subjects who are satisfied with the binary rule of enjoying local life is less than 50% and from the results of HAIM, we can conclude that increasing the opportunity to enjoy local life can improve tourist satisfaction effectively.

Table 4 Supports of variables selected in HAIM model

Rule	Percentage
I(Q1.Enjoyment of Local Life>3)	47%
I(Q2.Verbal Introduction to Traditional Life>1)	85%



10-fold CV for HAIM (a) 10-fold CV for AIM (b)

Fig. 6 The result of 10-fold-cross validation, (a) is the result of HAIM and (b) is the result of AIM. The horizontal axis shows the number of indices of HAIM and AIM. The vertical axis shows the corresponding means of score test statistics

Here, we also apply CART to the data, and the result of CART is shown in Fig. 7. We can confirm that the rules consist of "Q1 Enjoyment of Local Life", "Q8. Traditional Street Landscapes (Overall Atmosphere)", "Q17. Leisure and Rest" and "Q18. Enjoyment of Mahjong". We can see nearly half of the people (46%) are satisfied with Q1(Q1 $>$ 3) and almost all of them (90%) are satisfied with the tourist spot and both HAIM, AIM and CART select this rule. The result of CART is good at interpreting the interactions among selected variables. However, it cannot use a binary rule to interpret the result based only on one variable and the interpretation of the CART may be more complicated than those of HAIM and AIM.

Finally, we calculate the concordance index of HAIM, AIM and CART. The concordance index of HAIM (0.83) is higher than that of AIM (0.79) and CART (0.77). It shows that HAIM is well fitted to the data than AIM and CART. Additionally, the correct classification rate of HAIM (0.78), AIM (0.77) and CART (0.82) is also calculated.

6 Conclusion

In this paper, we proposed HAIM for binary response to obtain the score based only on controllable variables with the effects of uncontrollable variables are adjusted. In addition, we conducted numerical studies to compare the accuracy of HAIM to that of the other methods, including AIM and CART. The results show that HAIM have

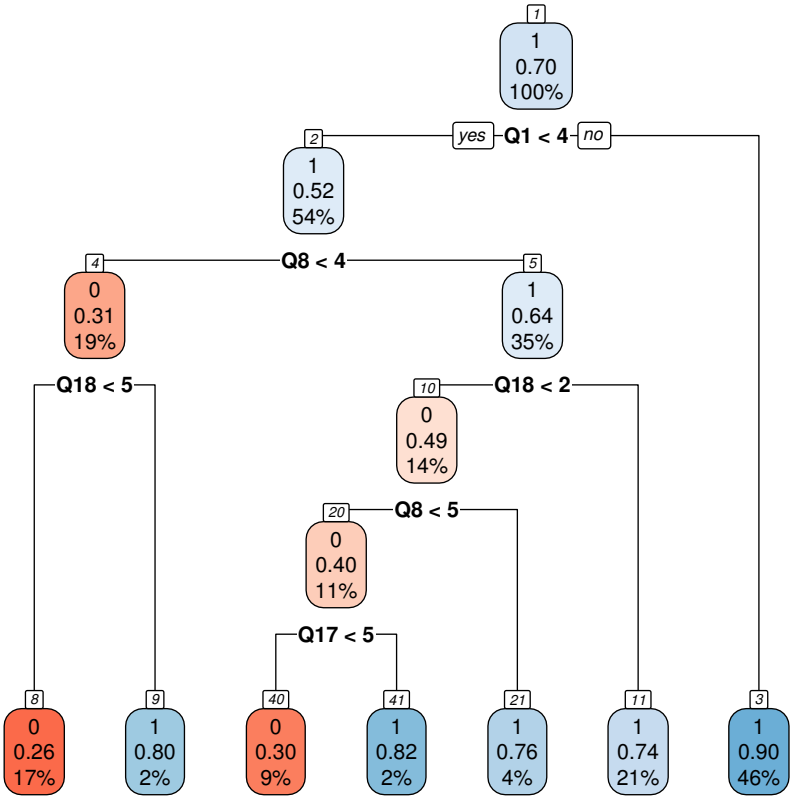


Fig. 7 The result of CART, blue boxes show the people who is satisfied with the tourist spot, red boxes show the people who is unsatisfied with the tourist spot. The middle of box shows the ratio of satisfaction(blue) or unsatisfaction(red) of tourist spot. The bottom of box shows the percentage of tourists, who satisfied the rules.

two features. First, the prediction accuracy of HAIM tends to be higher when the structure of data tends to be linear. Second, the correlation within the explanatory variables can improve the correct classification rate of HAIM when the structure of data tends to be linear. The application of HAIM in the real example also shows that HAIM is more appropriate than AIM in that situation. Also, the indices based only on controllable variables are constructed and revealed how the controllable variables affect tourist satisfaction. Furthermore, we explain the HAIM for binary response by using logistic model to adjust the effect of uncontrollable variables. It can also extend to survival response by using the cox regression to adjust the effect of uncontrollable variables and construct HAIM in the same algorithm.

Our results show that HAIM has several advantages, but there are still several future works. First, we must evaluate the results of the coefficients through testing.

For example, the power, type I error, and bias must be evaluated. Second, effects of variable selection for uncontrollable variables must be evaluated. Third, HAIM is combined model for AIM and linear model. It means that HAIM is quite straightforward and we need to consider the nonlinear effect of uncontrollable variables. In addition to that the effects of other kinds of response in HAIM should be considered.

Conflict of interest

We have no conflicts of interest to disclose

References

1. Breiman, L., Friedman, J.H., Stone, C.J., Olshen, R.A. *Classification and Regression Trees (1 edition edn)*. Chapman and Hall/CRC (1984)
2. Chen, G., Zhong, H., Belousov, A. and Deveanarayan, V. A PRIM approach to predictive-signature development for patient stratification. *Statist. Med.* 34, 317-342 (2015)
3. Friedman, J.H. Multivariate adaptive regression splines. *The Annals of Statistics.* 19, 1-67 (1991)
4. Friedman, J.H., Popescu, B.E. Predictive learning via rule ensembles. *Ann. Appl. Stat.* 2, 916-954 (2008)
5. Huang, X., Sun, Y., Trow, P., Chatterjee, S., Chakravartty, A., Tian, L. and Devanarayan, V. Patient subgroup identification for clinical drug development. *Statist. Med.* 36, 1414-1428 (2017)
6. Li, J., Jin, B. Multi-threshold Accelerated Failure Time Model. Annals of Statistics. 46: 2657-2682 (2018)
7. Lipkovich, I., Dmitrienko, A., Denne, J. & Enas, G. Subgroup identification based on differential effect search –a recursive partitioning method for establishing response to treatment in patient subpopulations, *Stat. Med.*, **30**, 2601–2621 (2011)
8. Liu, Y., Wan, K., Shimokawa, T. and Oyama, I. A Study on Current Situation of Tourism Development in Chengdu and the Suburb Area -Investigation of Affecting Factors of Tourist for Traditional Tourism Area in China-. *Transactions of Japan Society of Kansei Engineering*. 15 (1), 163-172 (2016)
9. Shimokawa, T., Oyama, I., Nishiyama, S., Kazama, F. and Kitamura, S. Development of the Elasticnet Regulation Boosting Trees for Binary Response: Application to the Questionnarie Data about Satisfaction of city water. *Transactions of Japan Society of Kansei Engineering*. 9 (4), 653- 661 (2009)
10. Shimokawa, T., Tsuji, M. and Goto, M. Extended Adaptive Index Models and its Evaluation. *Bulletin of Data Analysis of Japanese Classification Society*. 3 (1), 1-16 (2013)
11. Su, X., Meneses, K., McNees, P. and Johnson, W. Interaction trees: exploring the differential effects of an intervention programme for breast cancer survivors. *Appl. Statist.* 60 (3), 457-474 (2011)
12. Su, X., Tsai, C-L., Wang, H., Nickerson, D.M. and Li, B. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research.* 10, 141158 (2009)
13. Tian, L. and Tibshirani, R. Adaptive index models for marker-based risk stratification. *Biosatatistics.* 12(1), 68-86 (2010)
14. Wang, J., Li, J., Li, Y., Wong, W.K. A model-based multi-threshold method for subgroup identification. Statistics in Medicine. 38(14): 2605-2631 (2019)
15. Yan, K., Li, L., Shimokawa, T. and Kitamura, S. Exploration of factors Influencing on Evalution of Streetscape: A study on Cognitive Structure of Streetscape in Chendu: Part 2. *Transactions of Japan Society of Kansei Engineering*. 11 (1), 27- 37 (2012)