ES-DoS: Exhaustive search and density-of-states estimation as a general framework for sparse variable selection

Yasuhiko Igarashi\textsuperscript{1,2,3}, Hiroko Ichikawa\textsuperscript{4}, Yoshinori Nakanishi-Ohno\textsuperscript{1,5}, Hikaru Takenaka\textsuperscript{2}, Daiki Kawabata\textsuperscript{2}, Satoshi Eifuku\textsuperscript{6}, Ryoi Tamura\textsuperscript{7}, Kenji Nagata\textsuperscript{1,8} and Masato Okada\textsuperscript{2,3}

\textsuperscript{1}PRESTO, JST, Honcho 4-1-8, Kawaguchi, Saitama, 332-0012, Japan
\textsuperscript{2}Graduate School of Frontier Sciences, The University of Tokyo, Kashiwanoha 5-1-5, Kashiwa, Chiba, 277-8561, Japan
\textsuperscript{3}Research and Service Division of Materials Data and Integrated System, National Institute for Materials Science, Sengen 1-2-1, Tsukuba, Ibaraki, 305-0047, Japan.
\textsuperscript{4}Liberal Arts, Faculty of Science and Technology, Tokyo University of Science, Yamazaki 2641, Noda, Chiba, 278-8510, Japan
\textsuperscript{5}Graduate School of Arts and Sciences, The University of Tokyo, Komaba 3-8-1, Meguro, Tokyo 153-8902, Japan
\textsuperscript{6}Department of Systems Neuroscience, School of Science, Fukushima Medical University, Hikarigaoka, Fukushima City, Fukushima 960-1295, Japan
\textsuperscript{7}Graduate School of Medicine and Pharmaceutical Sciences, University of Toyama, Sugitani 2630, Toyama, Toyama, 930-0194, Japan
\textsuperscript{8}Artificial Intelligence Research Center, AIST, Aomi 2-3-26, Koto, Tokyo, 135-0064, Japan

E-mail: okada@k.u-tokyo.ac.jp

Abstract. In this paper, we propose an exhaustive search with density-of-states estimation (ES-DoS) method for sparse variable selection in a wide range of learning tasks with various learning machines. We applied this ES-DoS method to synthetic and real data as an example of the regression and classification problems and discuss the results in this paper. The most important aspect of our ES-DoS method is to extract not only the optimal solution but also density of states (DoS) in terms of machine learning and data-driven science. Mapping the solutions of various approximate methods or scientists' hypotheses onto the DoS, we can comprehensively discuss and evaluate these methods and hypotheses. Our ES-DoS method opens the way for sparse variable selection in various fields, which promotes the high-dimensional data-driven science.

1. Introduction
It is important to find effective explanatory variables for every machine-learning task, and this process is called variable selection, also known as subset selection [2–4]. What is essentially needed in variable selection is, besides what criterion, such as cross validation error, should be used, exhaustive search (ES) in which all possible sets of explanatory variables are evaluated and compared with one another in terms of a given criterion [5]. The history of ES in variable selection can be traced back a half a century [5,6], though it was mentioned in an early study [7] that ES was not necessary or desirable due to the proposal of stepwise regression algorithms [8].
Since then, it has been well known that in the variable selection problem, the number of all possible variable subsets is $2^N$, taking into account whether each variable is selected. The binomial theorem gives $2^N = \sum_{k=0}^{N} \binom{N}{k}$. This means that searching all the states naively in increasing or decreasing order of the number of explanatory variables used in each subset never reduces the computational complexity. Under such circumstances, in the linear-regression task, some reduction was accomplished by sharing part of the evaluation process between neighboring subsets and by applying search-tree techniques for efficient evaluation [9]. More recently, stochastic search methods, such as Markov chain Monte Carlo (MCMC, [10]), have been applied [11–14] in association with the development of Bayesian approaches [15–17]. From the viewpoint of sparse modeling, it is worthwhile to mention that an algorithm searching all possible subsets, the number of whose elements is less than or equal to a certain integer $K$, was proposed [18]. The motivation with this algorithm is closely related to that with a recently proposed method called $K$-sparse ES (ES-$K$), which searches all possible subsets that have exactly $K$ variables [19].

In this study, although previous studies have adopted ES only to search for the best subset of variables, it is argued that the ‘density of states (DoS)’ should be keenly appreciated to make further progress in ES. To emphasize the difference, we call our method as an exhaustive search (ES) and density of states (DoS) estimation (ES-DoS) method. The DoS represents a frequency distribution of criterion values taken by all the subsets, where a ‘state’ corresponds to a subset based on the analogy with statistical mechanics. Obviously, the DoS can be obtained by carrying out ES, but it is often intractable in the case of a large number of explanatory variables. Therefore, we previously developed an efficient method called the approximate ES (AES) method for estimating the DoS while avoiding a combinatorial explosion [1,20]. The AES method uses the multiple-histogram method [21] in combination with the exchange Monte Carlo (EMC) method [22], also known as parallel tempering [23]. One often falls into utter confusion when approximate algorithms, such as relaxation and greedy, provide different results. This might be caused by insufficient data in the first place. We discuss case studies to maintain that such problems can be solved by ES-DoS method for sparse variable selection.

The rest of this paper is organized as follows: Section 2 formulates ES-DoS method to be applicable to sparse variable selection in machine learning. In Section 3, as an example of the classification problem, we discuss applying the ES-logistic regression (ES-LoR) and ES-support vector machine (ES-SVM) method to synthetic and real data, respectively. In Section 4, we also discuss applying the ES-linear regression (ES-LiR) method to synthetic data as an example of the regression problem. Finally, the topics covered in the preceding sections are summarized, and the prospects of data-driven science and ES method are discussed.

2. Formulation

2.1. Machine learning

This section formulates ES-DoS method to make it available to a wide range of learning tasks with various learning machines, as shown in Fig. 1. The formulation accounts for supervised learning, such as regression and classification, but a similar formulation is considered applicable to unsupervised learning such as clustering with Gaussian mixture models.

In general, machine learning, such as regression and classification, is to make a learning machine $f(x; w)$ learn the relation between input $x = (x_1, x_2, \ldots, x_N)^T$ and output $y$ as follows:

$$y = f(x; w).$$  

(1)

Each component of input $x_i$ and that of output $y$ is called an explanatory variable and an objective variable, respectively. The variable $w$ denotes a set of parameters of the learning machine. As shown in the middle of Fig. 1, there are various learning machines, and which learning machine is used should be determined appropriately in accordance with each objective
Figure 1. Constitution of exhaustive search. Exhaustive search is conducted on every triplet of learning task, learning machine and criterion for variable selection.

Figure 1. Constitution of exhaustive search. Exhaustive search is conducted on every triplet of learning task, learning machine and criterion for variable selection.

task and computational theory [1]. For example, a linear model and logistic model are used in regression and classification tasks, and the coefficients $\mathbf{w} = (w_1, w_2, \ldots, w_N)^T$ are the parameters of the models as follows:

$$f(x; \mathbf{w}) = \sum_i w_i x_i,$$
$$f(x; \mathbf{w}) = 1/(1 + \exp(-\sum_i w_i x_i)).$$

When a Gaussian process is used as a learning machine, the parameters $\mathbf{w}$ characterize a kernel function [24]. In the case of a neural network, the parameters $\mathbf{w}$ represent weights.

The learning is, given $p$ pairs of learning data,

$$\mathbf{X} = \{x_{\mu}\}_{\mu=1}^p,$$
$$\mathbf{Y} = \{y_{\mu}\}_{\mu=1}^p,$$

defined as finding parameters $\mathbf{w}$ in machine learning to minimize a loss function. An example of a loss function is the following squared loss often used in regression tasks:

$$\epsilon(\mathbf{w}; \mathbf{X}, \mathbf{Y}) = \frac{1}{p} \sum_{\mu=1}^p (y_{\mu} - f(x_{\mu}; \mathbf{w}))^2.$$

In classification tasks, the logistic loss

$$\epsilon(\mathbf{w}; \mathbf{X}, \mathbf{Y}) = \frac{1}{p} \sum_{\mu=1}^p \log(1 + \exp(-y_{\mu} f(x_{\mu}; \mathbf{w})),$$

is often used, where $y_{\mu} \in \{-1, 1\}$.

2.2. State indicator

The problem of variable selection is discussed using an indicator $\mathbf{c} = (c_1, c_2, \ldots, c_N)^T \in \{0, 1\}^N$ to represent a set of explanatory variables [5, 13], where $c_i = 1$ indicates that the $i$-the explanatory variable $x_i$ belongs to the set and $c_i = 0$ indicates that it does not. In addition,
each realization of the indicator \( c \) is referred to as a ‘state’ based on the analogy with statistical mechanics. In the cases of a linear model and logistic model, the input-output relation described by the learning machine in state \( c \) is written as

\[
y = f(x \circ c, w \circ c),
\]

where the symbol \( \circ \) is the Hadamard product defined as \( (x \circ c)_i = x_i c_i \). This formulation can be generalized to every learning machine as follows:

\[
y = f_c(x(c); w(c)),
\]

where \( x(c) \) and \( f_c(x(c); w(c)) \) represent the selected variables and corresponding learning machine in state \( c \), respectively, and \( w(c) \) is a subset of \( w \) omitting all irrelevant components to \( x(c) \).

### 2.3. Criterion for variable selection

Although various criteria for variable selection have been proposed, as shown on the right of Fig. 1, we discuss the use of cross validation error (CVE) and Bayesian free energy (BFE) among the most popular criteria. Cross validation error is used to evaluate the predictability of selected explanatory variables and is asymptotically equivalent to the Akaike information criterion (AIC, [25]) in regular learning machines [26]. Bayesian free energy is defined as the negative logarithm of Bayesian posterior distribution, from which the Bayesian information criterion (BIC) is derived by approximation [27].

#### 2.3.1. Cross validation error

How to calculate CVE with respect to each state \( c \) is explained. The CV procedure is composed of three steps. First, the entire set of learning data \((X, Y)\) is divided into training data \((X_{tr}, Y_{tr})\) and test data \((X_{te}, Y_{te})\). Second, the learning machine \( f_c(X(c); w(c)) \) is optimized with respect to \( w(c) \) using only training data. The parameter set optimized at this training step is written as \( \hat{w}_{tr}(c) \). Finally, how accurately the trained machine can describe the input-output relation of test data is evaluated using a loss function, such as Eqs. (6) and (7), to obtain CVE:

\[
\text{CVE}(c) = \epsilon(\hat{w}_{tr}(c); X_{te}(c), Y_{te}).
\]

The CVE value depends on the method of dividing data, and CV is often conducted repeatedly with different partitions of data to obtain a typical CVE value. For example, in \( M \)-fold CV, all data are randomly divided into \( M \) subsets of the same size, and the mean of \( M \) CVE values is calculated by carrying out CV \( M \) times using each subset as validation data only once.

#### 2.3.2. Bayesian free energy

A conditional probability of the indicator \( c \) given learning data \((X, Y)\), which is called a posterior distribution, \( P(c|X, Y) \), is derived from the joint probability of learning data, parameters of a learning machine, and state indicator,

\[
P(X, Y, w(c), c).
\]

The parameters \( w(c) \) are marginalized out to obtain

\[
P(X, Y, c) = \int dw(c)P(X, Y, w(c), c).
\]

The product rule of probability is applied to the left-hand side to obtain

\[
P(c|X, Y)P(X, Y) = \int dw(c)P(X, Y, w(c), c).
\]
Again, the product rule is applied to the joint probability in the right-hand side to obtain

$$P(c|X,Y)P(X,Y) = \int dw(c) P(Y|X,w(c),c)P(X|w(c)|c)P(c),$$

where $P(w(c)|c)$ represents in a probabilistic way the fact that $w(c)$ is a subset of $w$ omitting all irrelevant components to $x(c)$, and $P(Y|X,w(c),c)$ represents the input-output relation, Eq. (9), with an appropriate noise model. It should be noted that $X$ is independent of $c$ and $w(c)$. By setting the prior distribution of $c$, $P(c)$ to the uniform distribution for simplicity, we obtain

$$P(c|X,Y) \propto \int dw(c) P(Y|X,w(c),c)P(w(c)|c).$$

It should be noted that $P(X,Y)$ and $P(X)$ are constant with respect to $c$.

Bayesian free energy is defined as the negative logarithm of a posterior distribution,

$$\text{BFE}(c) = -\log \int dw(c) P(Y|X,w(c),c)P(w(c)|c),$$

up to a constant. The formula of BFE is similar to that of free energy in statistical mechanics in that both have an integral with respect to many variables. This is because BFE is called a ‘Bayesian free energy’.

### 2.4. Exhaustive search with density-of-states estimation

The discussion in Subsections 2.1–2.3 are summarized as follows: When a learning machine is made to learn an input-output relation, such as Eq. (1), using $p$ pairs of input and output data $(X,Y)$, variable selection is carried out by evaluating each $c$ in a criterion such as CVE and BFE. Therefore, ES compares all the $2^N$ states to find the optimal one without fail.

Previous studies on ES were aimed only at the best subset of explanatory variables, namely, a ‘ground state’. However, it does not necessarily mean that they have brought out the full potential of ES. In this study, therefore, we argue that the DoS should be appreciated to make further progress in ES. When we emphasize the DoS, we especially call the method as an exhaustive search (ES) and density of states (DoS) estimation (ES-DoS) method. The DoS, a concept borrowed from statistical mechanics, represents a frequency distribution of criterion values within all the states. More formally, the DoS for each criterion $H$, such as CVE and BFE, is defined as follows:

$$D_H(E) = \sum_c 1(E = H(c)),$$

where the function $1$ takes the value 1 if the argument is satisfied or 0 if not. By definition, the DoS is successfully obtained with ES.

We previously proposed an approximate method for estimating the DoS with respect to CVE and BFE to circumvent a combinatorial explosion involved in ES due to a number of explanatory variables, and later named the method the approximate ES (AES) method [1, 20]. The AES method applies the multiple-histogram method [21] to histograms composed of state samples generated from probability distributions

$$p_\beta(c) \propto e^{-\beta H(c)}.$$

Based on the analogy with statistical mechanics, each of the distributions is called a ‘canonical distribution’ with a corresponding ‘temperature’ $T = \beta^{-1}$, and $H$, which represents a criterion,
is regarded as a `Hamiltonian'. It is convenient to use the EMC method for sampling from canonical distributions with different temperatures [22]. Similarly, an approximate version of the aforementioned ES-K method, i.e., AES-K, was developed [19].

In practical situations, one often comes across the following problem. When approximate algorithms, such as convex relaxation and greedy methods, give different solution states to the same task, it is confusing as to which should be trusted, and to make matters worse, it might be caused by insufficient data in the first place. As will be introduced in what follows, we have tried to solve such problems with close observation of the DoS.

![Figure 2. 4-dimensional synthetic data for classification](image)

3. Classification

3.1. ES-SVM and ES-LoR

In this section, we discuss the ES method for classification shown in Fig. 1. Let us start the discussion about the importance of variable selection by applying the ES method to the synthetic data illustrated in Fig. 2. Figure 2 represents the task of binary classification in that the elements have either +1 or -1. We discuss the task from the aspect of the linear classification.

The data represented in Fig. 2(a) are discriminated with both variables 1 and 2 clearly better than only with variable 1 or only with variable 2. On the other hand, in the case of Fig. 2(b), the data are discriminated with only variable 3 better than with both variables 3 and 4 because variable 4 is irrelevant for classification. In the case of Fig. 2(a), discriminative performance improves by using the two variables, while in the case of Fig. 2(b), discriminative performance improves by using only one variable. Although these can be easily found by the distribution of two variables, they cannot be found by the marginal distribution of each variable because the marginal distributions are not different. This means that the marginal distribution does not work for classification and that we have to search the multivariate distribution for classification, which causes an exponential increase in search space in the case of high dimensional input. This difficulty is solved using our ES method. In this section, we consider a support vector machine (SVM) as a machine-learning algorithm for classification and introduce ES for SVM (ES-SVM). The discriminative function and error function of an SVM are used based on a linear SVM as
follows,

\[ y = f_c(x(c); w(c)) = f(x; w \circ c) = \text{sgn}(w_0 + \sum_{i=1}^{4} c_i w_i x_i), \quad (19) \]

\[ \epsilon(w; Y, X) = \frac{1}{p_{te}} \sum_{\mu=1}^{p_{te}} \max(0, 1 - y_\mu(w_0 + \sum_{i=1}^{4} c_i w_i x_{i,\mu})) + \lambda \sum_{i=1}^{4} w_i^2. \quad (20) \]

For the criteria for variable selection, we used \( \text{CVE}(c) \), which are derived from 10-fold CV as follows:

\[ \text{CVE}(c) = \frac{1}{p_{te}} \sum_{\mu \in \text{te}} 1(y_\mu(w_0 + \sum_{i=1}^{4} c_i w_i x_{i,\mu}) < 0), \quad (21) \]

which represents the incorrect identification for test data, where \( p_{te} \) is the number of test data sets. We fixed \( \lambda = 5.0 \) in Eq. (20). Figure 3 shows the CVE for synthetic data obtained with the ES-1, -2, -3, and -4 methods. Among all the combination of variables (\( 2^4 - 1 = 15 \) combinations), the combination consisting of \( \{1, 2, 3\} \) had the lowest CVE, which is coincident with the above discussion. Since the combinations including variable 4 had a higher CVE than those excluding variable 4, this variable reduces the discriminative performance as noise. Cross validation error represents prediction error and is not appropriate for model selection. However, depending on the circumstance, we can appropriately select variables by CVE.

Similarly, we can consider logistic regression as a machine-learning algorithm for classification and introduce ES-LoR. We used Eqs. (3) and (7) as the discriminative function and error function of logistic regression, respectively. The results are shown in Fig. 3(b), where we used \( \text{CVE}(c) \) derived from Eq. (21) for the criteria for variable selection. Similar to the results of ES-SVM, the combination consisting of \( \{1, 2, 3\} \) had the lowest CVE among all the variable combination.

### 3.2. Application of the ES-SVM to the real data

We applied ES-SVM for the real data obtained from the neurophysiological experiment with macaque monkeys conducted by Eifuku et al. [20, 28, 29]. In this experiment, Eifuku et al. recorded the neuronal activities of 23 neurons in the anterior inferior temporal cortex (AIT) of macaque monkeys when the monkeys were performing a pattern-recognition task requiring...
the identification of facial images. The monkeys observed the 28 face images consisting of 4 different facial identities recorded from 7 different view and required to identify the same person irrespective of different facial views. We calculated the firing rate of 23 neuronal activities for each face image and investigated which neuron contributed to view-invariant face identification. It is a task of binary classification in that two out of four facial identities should be discriminated using the neuronal firing rate as the input data. We can pose the problems of linear discrimination among four identities; thus, the number of possible problems, that is the pairs of identities that should be discriminated, are $3C_2 = 6$. Hereafter, when we consider the problem of discrimination between identities nos. 1 and 3, the problem will be noted as “the discrimination of identity 1 vs. 3.”

In this study, we investigated the variable selection in the real data by using the weight diagram and DoS, focusing on the discrimination of identities 1 vs. 3 and 1 vs. 4. They were investigated by applying the ES-SVM by Nagata et al. [20]. For the ES-SVM applied to 23 explanatory variables, the SVM classifications and CVE calculations by leave-one-out cross validation (LOOCV) were conducted for the $2^{23} - 1 = 8,388,607$ possible variable subsets.

3.2.1. Analysis of the weight diagram and density of states (DoS) The weight diagram illustrates the weight coefficients of each explanatory variable $w(c)$ for calculating the decision boundary. Using the weight diagram, we visualized the weight coefficients $w(c_{opt})$ of each explanatory variable determined by the optimal subsets $c_{opt}$ with CVE = 0.

Figure 4 shows the weight diagram of optimal subsets for the discrimination of identities (a)
Figure 5. DoS for discrimination of identity 1 vs. 3. Each figure shows DoS for $K = 1, \ldots, 6$. Vertical axes indicate frequency and horizontal axes indicate CVE.

1 vs .3 and (b) 1 vs 4. In each weight diagram, the vertical axis is the vector indicating the index of the explanatory variable (e.g. neuron index) with length of $|x|$. The horizontal axis is the vector indicating the index of the optimal subset. Since the ES-SVM for the discrimination of identity 1 vs. 3 found 166,412 optimal subsets, the length of the horizontal axis in Fig. 4 (a) is 166,412. Similarly, the ES-SVM for the discrimination of identity 1 vs. 4 found 1,938 optimal subsets, and the length of the horizontal axis in Fig. 4 (b) is 1,938. A column in the weight diagram indicates an optimal subset $c_{opt}$. In each column, a colored cell indicates that the explanatory variable consists of the optimal subset and its indicator $c_i = 1$. The color of the cell indicates a weight coefficient $w(c_{opt})$ for calculating the decision boundary. The red/yellow gradient indicates positive value and blue gradient indicates negative value. In each column, the non-colored cell, i.e., black cell, indicates that the explanatory variable was not selected as the optimal subset and $c_i = 0$. As shown in Fig. 4, the weight diagram enabled us to capture the whole picture of optimal subsets that cannot be provided by seeking one of the optimal subsets.

As shown in a previous study [20], we compared the DoS of the CVE obtained from the ES-SVM with DoS for random guessing. It is represented by binomial distribution DoS, where the given data $y$ were randomly labeled regardless of input data $x$. Figure 4 (c) shows the DoS obtained from the ES-SVM on the discrimination of identity 1 vs. 3. Compared with the DoS for random guessing, the DoS is asymmetric and left-skewed distribution. This result suggests that ES-SVM successfully classified the neuronal activity data into identity 1 or 3 more frequently rather than the chance level, and our data include significant information for binary classification of identity 1 vs. 3. On the other hand, Fig. 4(d) shows the DoS obtained from the ES-SVM on the discrimination of identity 1 vs. 4. Similar to the DoS for random guessing, the DoS is a symmetric distribution, Nagata et al. [20] discussed that ES-SVM successfully classified the neuronal activity data into identity 1 or 4 at the chance level, and the data did not include significant information for binary classification of identity 1 vs. 4.

In the present study, we applied ES-$K$ to the discrimination of identities 1 vs. 3 and 1 vs. 4.
In order of increasing $K$, we investigated the structure of optimal subset $e_{opt}$, which consists of $K$ explanatory variables. First, we used ES-1 for the discrimination of identity 1 vs. 3. Figure 5(a) shows the DoS obtained from ES-1. We did not find the optimal subset with CVE = 0. Figure 5(b) shows the DoS obtained from ES-2. We found the two optimal subsets, neurons nos. 6 and 13 and neurons nos. 16 and 20. Hereafter, we note the variable indices (that is, neuron index) included in the subset as $\{6, 13\}$ and $\{16, 20\}$. As shown in Fig. 5(c), we found the 16 optimal subsets from ES-3. Out of these optimal subsets, 11 subsets contained $\{6, 13\}$ and 3 contained $\{16, 13\}$. In other words, these 14 subsets were supersets of the optimal subsets found with ES-2. When the optimal subsets found with ES-$K$ are supersets of the optimal subsets found with ES-$(K - 1)$, we called the relation the “parent-children” relation.

Figure 6(a) illustrates an overview of the “parent-children” relation. We call the two optimal subsets found with ES-2, $\{6, 13\}$ and $\{16, 20\}$, as the ES-2 “parents”. We also call the 14 optimal subsets found with ES-3, $\{6, 13, \alpha\}$ and $\{16, 20, \beta\}$ as ES-2’s “children”, where $\alpha$ and $\beta$ indicates the arbitrary neuron index. The case in which $K \geq 3$, we call the optimal subsets found with ES-$K$, which are not ES-$(K - 1)$’s children as ES-$K$’s “parents”. For example,
for the above discrimination of identity 1 vs. 3, the two optimal subset found with ES-3, 
\{13, 16, 21\} and \{13, 16, 23\}, are ES-3 \( \sqsubseteq \) “parents”. Therefore, the optimal subsets found with 
ES-\( K \) are composed of ES-(\( K - 1 \))’s “children” and ES-\( K \) \( \sqsubseteq \) “parents”. The proportion of 
ES-(\( K - 1 \))’s “children” and ES-\( K \) \( \sqsubseteq \) “parents” obtained with ES-\( K \) at each \( K \) is shown in 
Fig. 6(c). Irrespective of the increment of \( K \), most of the optimal subsets were ES-(\( K - 1 \))’s 
“children”. This result indicates that the optimal subsets found by relatively small \( K \) (here, 
\( K = 2 \)) can be described as “parents” and most of the optimal subsets found by relatively large \( K \) 
(here, \( K \geq 3 \)) consist of the “parents”. Since the optimal subsets commonly contain a smaller 
number of variables as “parents”, it is important for feature selection to find the “parents”. 
Next, we applied ES-\( K \) to the discrimination of identity 1 vs. 4. In this case, we could not 
find an optimal subset in the case of \( K = 1, 2, 3 \). As Fig. 6(b) illustrates, we found the three 
optimal subsets from ES-4. In contrast to the discrimination of identity 1 vs. 3, along with 
the increment of \( K \), the number of optimal subsets found with ES-\( K \) increased. We found the 
5 optimal subsets from ES-5 and 17 optimal subsets from ES-6. Therefore, the proportions of 
ES-(\( K - 1 \))’s “children” in the optimal subset found in each \( K \) were low at each \( K \), 
as shown in Figure 6(d). Finally, the proportion of ES-(\( K - 1 \))’s “children” among the 1,936 
optimal subsets obtained with ES-SVM (that is, ES-K-SVM conducted every \( K \) from 1 to 23) 
was about 50\%. This result indicates that the optimal subsets for the discrimination of identity 1 
vs. 4 did not consist of the “parent-children” relation. Since the “parents” that constantly took 
part in the multiple optimal subsets are not stable, it is difficult to determine which exploratory 
variable should be selected.

In summary, in the discrimination of identity 1 vs.3, the optimal subsets are related to each 
other in the ‘parent-children” relation, and the DoS distribution was a left-skewed distribution 
and differed from the DoS for random guessing. In the discrimination of identity 1 vs.4, 
the optimal subsets did not construct the “parent-children” relation, and the DoS distribution 
overlapped with the DoS for random guessing. Therefore, the existence of the “parent-children” 
relation in the classification problem suggests that we can select the “parent” as an important 
explanatory variable for classification. This is closely related with the difference in the DoS 
distribution from the DoS for random guessing [20]. Based on these findings, our ES method 
provides insight into determining whether the data contain the information for classification. 
Furthermore, ES-\( K \) would help variable selection for classification.

4. Regression

4.1. Formulation of exhaustive search for linear regression

In this section, we consider regression as a machine-learning task and introduce exhaustive search 
for linear regression (ES-LiR). Let us specifically formulate ES-LiR. Using Eq. (9), we can write 
f_\epsilon(x(e); w(c)) as

\[ y = f_\epsilon(x(e); w(c)) = f(x; w \circ c) = w_0 + \sum_{i=1}^{N} c_i w_i x_i. \] (22)

We use CVE and BFE as the evaluation criteria of variable selection, as denoted in §2.3. Cross 
validation error is derived from CVE(e) using the error function \( \epsilon(w; Y, X) \) described 
in Eqs. (10) and (6), respectively. Similarly, we introduce BFE in Subsubsection 2.3.2 for linear 
regression. Using Eq. (22), we can write \( P(Y|X, w(c), c) \) from Eq. (16),

\[ P(Y|X, w(c), c) = \frac{1}{\det(2\pi \Sigma)^{1/2}} \exp \left( -\frac{1}{2} \Delta^{T} \Sigma^{-1} \Delta \right), \quad \Delta = [Y - \{w_0 + X(c \circ w)\}], \] (23)

where \( 1 \) is a \( p \)-dimensional all-one vector, and \( \Sigma \) represents the covariance matrix of measurement 
noise, whose elements are given by \( \Sigma_{ii} = \sigma^2 \) and \( \Sigma_{ij} = 0 \) (\( i \neq j \)).
Let us consider $P(\mathbf{w}(c)|c)$ in Eq. (16). We assume that $w_i$ in the case of $c_i = 1$ is generated from the Gaussian distribution where the mean and variance are 0 and $s^2$, respectively, and $w_0$ in Eq. (22) is also generated from the same distribution. Then, we can write BFE in Eq. (16) as

$$\text{BFE}(c) = -\log \int \frac{dw_0}{\sqrt{2\pi}s^2} \exp \left( -\frac{w_0^2}{2s^2} \right) \prod_{i|c_i=1} \frac{dw_i}{\sqrt{2\pi}s^2} \exp \left( -\frac{w_i^2}{2s^2} \right) P(Y|\mathbf{X}, \mathbf{w}(c), c) (24)$$

where we set $\Lambda = (\mathbf{X}_1^T\Sigma^{-1}\mathbf{X}_1 + \frac{1}{2}\mathbf{I})^{-1}$, $\mu = \Lambda\mathbf{X}_1^T\Sigma^{-1}\mathbf{y}$ and $\mathbf{X}_1$ is a matrix composed of non-zero explanatory variables. We estimate the variance $s^2$ using the observed data as described in the next paragraph. With our ES method, the Bayesian free energy (BFE) is calculated for all combinations of explanatory variables, and the combination minimizing the BFE is taken as the optimal one.

4.2. Results
We applied ES-LiR to a synthetic dataset for sparse variable selection in linear regression. We derived the DoS for BFE(c) and CVE(c) and investigated the solution space. For the synthetic data, we assumed that the $N$ dimensional coefficient vector $\mathbf{w}$ has two non-zero variables and set $w_1 = -0.65, w_2 = 1.2, w_i = 0 (i \neq 1, 2)$. The synthetic data $y_\mu$ were obtained as follows

$$y_\mu = w_0 + \sum_{i=1}^{N} w_i x_{i,\mu} + \epsilon_\mu. (26)$$

Each element of $x_{i,\mu} (\mu = 1, \ldots, p)$ is generated from the Gaussian distribution $\mathcal{N}(0, \sigma^2_w)$ where the mean and variance are 0 and $\sigma^2_w$, respectively. Each noise component $\epsilon_\mu$ is assumed to follow $\mathcal{N}(0, \sigma^2_\epsilon)$, and we set $\sigma^2_\epsilon = 0.1$, signal-to-noise ratio was set as $\sigma^2_w/\sigma^2_\epsilon = 10$, and the data dimension and sample number as $N = 200$ and $p = 50$, respectively.

4.2.1. DoS and solution analysis using ES-LiR We applied ES-LiR to the synthetic data and investigated whether the method can extract the non-zero variables $\{1, 2\}$ and visualize the two-dimensional DoS corresponding to CVE and BFE for studying the solution structure. Figure 7 shows the DoS for synthetic data obtained with ES-1, -2, and -3. The combination consisting of $\{1, 2\}$ had a remarkably lower CVE and BFE than the other combinations of two explanatory variables, as shown in Fig. 7(b). Then let us consider the results of ES-3. Figure 7(c) shows the structure consisting of combinations of two variables including $\{1, 2\}$ and another variable. By comparing CVE and BFE, we found that BFE leads to better understanding of the solution structure than CVE, as shown in Fig. 7. Similar results were obtained in recent research in which parts of the authors calculated the DoS for an astronomical dataset [19]. We can estimate the sparseness $K_{\text{opt}}$ from the data by calculating the minimum values of BFE and CVE for all combinations of $K = 1, 2, \ldots, N$ explanatory variables and comparing the values [19].

5. Discussion and conclusion
In this paper, we proposed an exhaustive search with density-of-states estimation (ES-DoS) method for sparse variable selection. We applied ES-linear regression (ES-LiR) to synthetic data as an example of the regression problem. As an example of the classification problem, we also applied ES-logistic regression (ES-LoR) and ES-support vector machine (ES-SVM) to synthetic and real data, respectively. Practically, to apply our ES method to data analysis, we
Figure 7. DoS for synthetic data obtained using ES-1, -2, -3. In each figure, horizontal and vertical histograms represent DoS corresponding to CVE and BFE, respectively, and central figure shows two-dimensional DoS. They are expressed in logarithmic scales. Combinations forming each cluster are written in \( \{ \} \).

used \( K \)-sparse ES (ES-\( K \)), which searches all possible subsets that have exactly \( K \) variables [19] and is important for understanding the solution structure, as shown in §3.2.

The most important aspects of our ES-DoS method is to extract not only the optimal solution [2] but also density of states (DoS). We discussed the importance of the DoS in terms of machine learning and data-driven science [1]. Generally, various approximate methods, such as LASSO [30], and Iterative thresholding [31], can be applied to the target measurement data. Using the state indicator of solutions obtained with these methods, we can map the solutions of various approximate methods onto the DoS. We thus can comprehensively discuss and evaluate the advantages and disadvantages of each method.

We then discussed the importance of the DoS for data-driven science [1]. Depending on the various scientific disciplines, each scientist forms a hypothesis and conducts experiments and measurements for hypothesis testing. Generally, the hypothesis is formulated as a feature selection to determine which variable essentially explains the data. We can map the scientist’s hypothesis onto the DoS using the state indicator corresponding to the hypothesis. We thus can comprehensively discuss and evaluate the different hypotheses. From this standpoint, the solution obtained with an approximate method is regarded as one of the hypotheses.

The proposed method, ES-DoS method, opens the way for sparse variable selection in various fields, which promotes the high-dimensional data-driven science.

Acknowledgments
This work was supported by a Grant-in-Aid for Scientific Research on Innovative Areas (No. 25120009, 16H01555), Grant-in-Aid for Young Scientists (B) (No. 17K12735,17K12749) and Grants-in-Aid for JSPS Fellows (No. 15J07765) from Japan Society for the Promotion of Science (JSPS) and by JST CREST(JPMJCR1761) and PRESTO (JPMJPR15E8, JPMJPR17N2, JPMJPR1773).

References
[1] Igarashi Y, Nagata K, Kuwataki T, Omori T, Nakanishi-Ohno Y, and Okada M 2016 J. Phys. Conf. Ser. 699 012001
[2] Beale E M L, Kendall M G, and Mann D W 1967 Biometrika 54 357–366
[3] Miller A J 1990 Subset selection in regression Chapman and Hall
[4] George E I 2000 *J. Am. Stat. Assoc.* **95** 1304–1308
[5] Garside M J 1965 *J. R. Stat. Soc. Ser. C* **14** 196–200
[6] Kudō A 1963 *Memoirs of the Faculty of Science, Kyushu University. Series A, Mathematics* **17** 63–75
[7] Healy M J R 1963 *Comput. J.* **6** 57–61
[8] Efroymson M A 1960 *Mathematical Methods for Digital Computers* John Wiley and Sons 191–203
[9] Furnival G M and Wilson R W 1974 *Technometrics* **16** 499–511
[10] Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H, and Teller E 1953 *J. Chem. Phys.* **21** 1087–1092
[11] George E I and McCulloch R E 1993 *J. Am. Stat. Assoc.* **88** 881–889
[12] Green P J 1995 *Biometrika* **82** 711–732
[13] Kuo L and Mallick B 1998 *Sankhyā Ser. B* **60** 65–81
[14] Kim S, Tadesse M G, and Vannucci M 2006 *Biometrika* **93** 877–893
[15] Lindley D V 1968 *J. R. Stat. Soc. Ser. B* **30** 31–66
[16] Mitchell T J and Beauchamp J J 1988 *J. Am. Stat. Assoc.* **83** 1023–1032
[17] Ishwaran H and Rao J S 2005 *Ann. Stat.* **33** 730–773
[18] Tarumi T and Kudo A 1974 *J. Japan. Statist. Soc.* **4** 47–56
[19] Igarashi Y, Takenaka H, Nakanishi-Oino Y, Uemura M, Ikeda S, and Okada M 2017 *arXiv preprint arXiv:1707.02050*
[20] Nagata K, Kitazono J, Nakajima S, Eifuku S, Tamura R, and Okada M 2015 *IPSJ Online Trans.* **8** 25–32
[21] Ferrenberg A M and Swendsen R H 1989 *Phys. Rev. Lett.* **63** 1195–1198
[22] Hukushima K and Nemoto K 1996 *J. Phys. Soc. Jpn.* **65** 1604–1608
[23] Geyer C J 1991 in *Proceedings 23th Symp. Interface Comput. Sci. Stat* 156–163
[24] Dam H C 2016 private communications
[25] Akaike H 1973 in *2nd Int. Symp. Inf. Theory* 267–281
[26] Stone M 1977 *J. R. Stat. Soc. Ser. B* **39** 44–47
[27] Schwarz G 1978 *Ann. Stat.* **6** 461–464
[28] Eifuku S, De Souza W C, Tamura R, Nisijio H, and Ono T 2004 *J. Neurophysiol* **91**, 358–371
[29] Eifuku S, De Souza W C, Nakata R, Ono T, and Tamura R 2011 *PLoS One* **6**, e18913
[30] Tibshirani R 1996 *J. Royal Stat. Soc. B* **58**(1), 267-288
[31] Blumensath T, and Mike E D 2008 *J. Fourier Analys and Applications* **14**(5), 629-654