IPSJ Transactions on Mathematical Modeling and Its Applications Vol.8 No.2 23–30 (July 2015)

Regular Paper

An Exhaustive Search and Stability of Sparse Estimation for Feature Selection Problem

Kenji Nagata[1,a)] Jun Kitazono[2] Shinichi Nakajima[3] Satoshi Eifuku[4] Ryoi Tamura[5]
Masato Okada[1,6,b)]

Received: August 25, 2014, Revised: October 17, 2014,
Accepted: December 24, 2014

Abstract: Feature selection problem has been widely used for various fields. In particular, the sparse estimation has the advantage that its computational cost is the polynomial order of the number of features. However, it has the problem that the obtained solution varies as the dataset has changed a little. The goal of this paper is to exhaustively search the solutions which minimize the generalization error for feature selection problem to investigate the problem of sparse estimation. We calculate the generalization errors for all combinations of features in order to get the histogram of generalization error by using the cross validation method. By using this histogram, we propose a method to verify whether the given data include information for binary classification by comparing the histogram of predictive error for random guessing. Moreover, we propose a statistical mechanical method in order to efficiently calculate the histogram of generalization error by the exchange Monte Carlo (EMC) method and the multiple histogram method. We apply our proposed method to the feature selection problem for selecting the relevant neurons for face identification.

Keywords: feature selection, exhaustive search, cross validation, exchange Monte Carlo method

1. Introduction

Feature selection problem, in which a combination of relevant features is selected from the given features, is an important process for improving the generalization capability and interpretability of learning models [6], [10]. Cover and Van Campenhout showed that no non-exhaustive sequential feature selection procedure can be guaranteed to produce the optimal subset in the feature selection problem [1]. However, the exhaustive search method requires huge computational cost because the number of possible subsets grows combinatorially as the dimension of data increases.

From this problem, the various algorithms of feature selection problem have been developed, which is reviewed in Ref. [10]. Sparse estimations such as Least absolute shrinkage and selection operator (LASSO) [12] and automatic relevance determination (ARD) [15] are one of these algorithms, and widely used in recent studies. The sparse estimation is a method to minimize the error function with a small number of features. One advantage of the sparse estimation is that its computational cost is the polynomial order of the number of features. Another advantage is that the algorithm can search a unique optimal solution because the algorithm of sparse estimation reduces to the convex optimization problem. On the contrary, the sparse estimation has the problem that the obtained solution varies as the dataset has changed a little [7], [11], [14]. In this case, the solution also varies as the algorithm of sparse estimation changes, and hence, it is difficult to select the relevant features.

In order to clarify the reason of this problem, we need to know the structure of solution space for training task. The goal of this paper is to exhaustively search the solutions which minimize the generalization error for feature selection problem. This paper focuses on the feature selection problem for the binary classification with linear discriminant. We use the support vector machine (SVM) as the linear discriminant, and the cross validation method for the estimation of the generalization error for a combination of features. We calculate the generalization errors for all combinations of features in order to get the histogram of generalization error. By using this histogram, we propose a method to verify whether the given data include information for binary classification by comparing the histogram of generalization error for random guessing. This proposed method gives us a new insight compared to the conventional method such as sparse estimation, in which we assume that the given data have information significantly for binary classification.

Moreover, we propose a statistical mechanical method in order to efficiently calculate the histogram of generalization error. Typically, it costs exponential order of the number of features to obtain the histogram of generalization error. Hence, it is necessary for developing the algorithm to efficiently calculate the histogram of the generalization error in the case of high-dimensional data analysis. In this paper, we propose the method for efficiently

1 The University of Tokyo, Kashiwa, Chiba 277–8561, Japan
2 Kobe University, Kobe, Hyogo 657–8501, Japan
3 Technische Universitat Berlin, German
4 Fukushima Medical University, Fukushima 960–1295, Japan
5 The University of Toyama, Toyama 930–0194, Japan
6 JST ERATO OKANOYA EMOTIONAL INFORMATION PROJECT, Wako, Saitama 351–0198, Japan
a) nagata@mns.k.u-tokyo.ac.jp
b) okada@k.u-tokyo.ac.jp

calculating the histogram of generalization error by combining the exchange Monte Carlo (EMC) method [8] and the multiple histogram method [9].

We apply our proposed method to the feature selection problem for selecting the relevant neurons for face identification in order to check the effectiveness of the proposed method. We show that, for identification of a certain pair of faces, the result for conventional sparse estimation varies so that we cannot recognize which neurons are relevant for face identification. Moreover, we show that the data for face identification of this pair do not include significant information by using the proposed verification method.

2. Method

In this section, we firstly describe feature selection problem for binary classification with linear discriminant. We propose a method to estimate the generalization performance for all combination of features, and to evaluate whether the training data given have information for desired binary classification by using the histogram of generalization error.

2.1 Binary Classification Problem with Linear Discriminant

Firstly, we formulate binary classification problem with linear discriminant.

The problem treated here is a binary classification problem using training data set;

$$\left\{(\boldsymbol{x}_i, t_i)|\boldsymbol{x}_i = \{x_{ij}\}_{j=1}^{D} \in \mathbb{R}^D, t_i \in \{+1, -1\}\right\}_{i=1}^{N}, \tag{1}$$

where $\boldsymbol{x}_i$ is a D-dimensional feature vector, t_i is a class label of $\boldsymbol{x}_i$, and N is the number of samples. Given the data set, the goal of binary classification is to find a hyperplane in the feature vector space that separates the samples with $t_i = 1$ from those with $t_i = -1$ by using this data set. The obtained hyperplane is referred to as a decision boundary, and is expressed by a linear equation:

$$y(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = 0, \tag{2}$$

where $\boldsymbol{w}$ is a weight vector. The goal is to find $\boldsymbol{w}$ and b that satisfy $y(\boldsymbol{x}_i) > 0$ for $t_i = 1$ and $y(\boldsymbol{x}_i) < 0$ for $t_i = -1$, that is, $ty(\boldsymbol{x}) > 0$ for all samples.

There are several learning machines for binary classification with linear discriminant. In this paper, we use support vector machine (SVM), which is described in Appendix in detail. SVM has been widely used as a learning machine with high generalization performance based on the concept of maximization of margin. However, SVM does not have an ability of feature selection because learning algorithm of SVM is constructed by using all of features given.

In this study, we use the cross validation in order to select relevant features from the given features. The cross validation (CV) is a model validation technique for estimating the generalization performance of unknown data. In the CV, the given data is divided into two parts. One part is used for the training of the model, and the other part is for validating the generalization performance of the model. This training and validating operation are iterated with different partitioning in order to reduce variability. In this study, we use leave-one-out CV (LOOCV), in which the number of test data is set one.

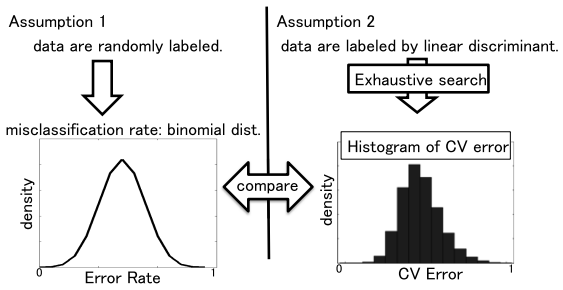

Fig. 1 Schematic figure of the proposed method.

Let us formulate concretely. We define an indicator variable $\boldsymbol{s} = \{s_j\}_{j=1}^{D} \in \{0, 1\}^D$, which represents a combination of features. The variable $\boldsymbol{s}$ indicates that $s_j = 1$ if the j-th feature is contained in the combination, and $s_j = -1$ if it is not. By using the combination of features represented by the indicator variable $\boldsymbol{s}$, we calculate the predictive error rate $E(\boldsymbol{s})$ for test data. When the predictive error rate $E(\boldsymbol{s})$ is small, the corresponding combination of features provides good generalization performance.

2.2 Exhaustive Search and Histogram of $E(\boldsymbol{s})$

In this study, we evaluate the generalization performance by calculating the predictive error rate $E(\boldsymbol{s})$ for all combination of features in order to clarify the solution-space structure of the predictive error rate $E(\boldsymbol{s})$. In particular, we propose a method to judge whether the given data include information for the desired binary classification by using the histogram $g(E)$ of the predictive error rate or not. When the histogram $g(E)$ has much density in small values of E, the given data can be considered to contain much information for binary classification.

We also propose a method to compare the obtained histogram $g(E)$ to the histogram $g_{bi}(E)$ for random guessing. **Figure 1** shows a schematic figure of the proposed method. More specifically, we assume that the given data are randomly labeled regardless of input data $\boldsymbol{x}$. In this case, the probability that each datum is correctly labeled was 0.5, and the error rate for N data is given by the following binomial distribution

$$p(k) = \frac{N!}{k!(N-k)!}(0.5)^k(1-0.5)^{N-k}, \tag{3}$$

where k is the number of misclassifications. From the distribution $p(k)$, we construct the histogram $g_{bi}(E)$ by plotting $E = k/N$ for horizontal axis and $g_{bi}(E) = p(k) \times (2^D - 1)$ for vertical axis. If the histogram $g(E)$ is significantly different from the histogram $g_{bi}(E)$ for random guessing, the given data have information for the desired binary classification. Otherwise, the data have no such information.

In order to quantitatively evaluate the difference between the two distribution $g(E)$ and $g_{bi}(E)$, we use the following Kullback divergence,

$$KL(g|g_{bi}) = \sum_{E} g(E) \log \frac{g(E)}{g_{bi}(E)} \tag{4}$$

The Kullback divergence has the property that $KL(g|g_{bi}) \geq 0$, and that $KL(g|g_{bi}) = 0$ if and only if $g(E) = g_{bi}(E)$. Hence, when the value of Kullback divergence is small, the given data have information for the desired binary classification.

2.3 Efficient Exhaustive Search Method

In order to carry out the method described in the above section, we require the histogram $g(E)$ of the predictive error rate, which spends the huge computational cost proportional to the exponential order of the dimension D. This is a terrible problem. In this section, we propose a statistical mechanical method, which consists of the exchange Monte Carlo method and Multiple histogram method, in order to overcome the problem of computational cost.

2.3.1 Exchange Monte Carlo Method

The exchange Monte Carlo (EMC) method is an algorithm of the Markov chain Monte Carlo (MCMC) method, and is to efficiently sample from the desired probability distribution. The purpose of the EMC method in this study is to sample from the following probability distribution with the CV error $E(\boldsymbol{s})$,

$$p(\boldsymbol{s};\beta) = \frac{1}{Z_\beta}\exp(-\beta E(\boldsymbol{s})), \tag{5}$$

where $\beta > 0$ is the parameter called "inverse temperature." The probability distribution $p(\boldsymbol{s};\beta)$ has high density in the state $\boldsymbol{s}$ with small CV error $E(\boldsymbol{s})$. Hence, the sampling from the probability distribution $p(\boldsymbol{s};\beta)$ corresponds to the search of the state $\boldsymbol{s}$ with small CV error $E(\boldsymbol{s})$.

In the EMC method, we prepare for replicas of the probability distribution $p(\boldsymbol{s};\beta)$ with several inverse temperatures $0 = \beta_1 < \beta_2 < \cdots < \beta_M$. That is to say, the EMC method is considered to sample from the following joint probability distribution $p(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M)$,

$$p(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M) = \prod_{m=1}^{M} p(\boldsymbol{s}_{\boldsymbol{m}};\beta_m). \tag{6}$$

The detailed algorithm of EMC method consists of the following two steps.

Step 1: update for each replica

Update the state $\boldsymbol{s}$ for each replica by the Metropolis algorithm, which is the most fundamental algorithm of MCMC method. Then, the state $\boldsymbol{s} = (s_1,\ldots,s_d,\ldots,s_D)$ is changed to the state $\boldsymbol{s}' = (s_1,\ldots,-s_d,\ldots,s_D)$ with the following probability,

$$p(\boldsymbol{s} \to \boldsymbol{s}') = \min\left(1, \exp(-\beta(E(\boldsymbol{s}') - E(\boldsymbol{s})))\right) \tag{7}$$

Step 2: exchange between the neighboring replicas

Exchange the states between the neighboring replicas, that is, $\{\boldsymbol{s}_m, \boldsymbol{s}_{m+1}\} \to \{\boldsymbol{s}_{m+1}, \boldsymbol{s}_m\}$, with the following probability,

$$\begin{aligned} p(\boldsymbol{s}_m \leftrightarrow \boldsymbol{s}_{m+1}) &= \min(1, v) \\ v &= \frac{p(\boldsymbol{s}_{m+1};\beta_m)p(\boldsymbol{s}_m;\beta_{m+1})}{p(\boldsymbol{s}_m;\beta_m)p(\boldsymbol{s}_{m+1};\beta_{m+1})} \\ &= \exp\left((\beta_{m+1}-\beta_m)(E(\boldsymbol{s}_{m+1}) - E(\boldsymbol{s}_m))\right). \end{aligned}$$

After iterating these two steps many times, the obtained distributions of states $\{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}$ converge to the joint probability distribution $\prod_{m=1}^{M} p(\boldsymbol{s}_{\boldsymbol{m}};\beta_m)$. Then, we obtain the samples from the probability distributions $p(\boldsymbol{s}_{\boldsymbol{m}};\beta_m)$ for each inverse temperature β_m.

One of the advantages for the EMC method is that this method can search the global optimal solution even though the error function has many local minima. As above-mentioned, the calculation of the error functions for all combinations of features requires huge computational cost. On the other hand, the method to search the solution locally such as the gradient descent method has the risk to trap the local minima. The EMC method enables us to search the global optimal solution efficiently because we can avoid that the sample traps the local minima by the exchange process in EMC method.

2.4 Multiple Histogram Method

Our proposed method cannot only search the combinations of features with minimal CV error, but also can estimate the histogram $g(E)$ of CV error by combining the multiple histogram method [4], [9]. The EMC method can search many optimal solutions efficiently. However, it is difficult to search all possible optimal solutions by the EMC method because this method is based on the probabilistic algorithm. For this problem, by combining this method to the multiple histogram method, we can estimate the histogram $g(E)$ of CV error by using sample sequence of states $\{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_M\}$ generated by the EMC method.

The histogram $g(E)$ is also called density of states, and is defined as follows,

$$g(E) = \sum_{\boldsymbol{s}} \delta\left(E - E(\boldsymbol{s})\right), \tag{8}$$

where $\delta(E)$ is the Dirac delta function. The density of states $g(E)$ has the relationship between the normalization constant Z_β for Eq. (5) as follows,

$$\begin{aligned} Z_\beta &= \sum_{\boldsymbol{s}} \exp\left(-\beta E(\boldsymbol{s})\right) && (9)\\ &= \sum_{\boldsymbol{s}}\sum_{E} \delta\left(E - E(\boldsymbol{s})\right)\exp\left(-\beta E(\boldsymbol{s})\right) && (10)\\ &= \sum_{E} g(E)\exp\left(-\beta E\right). && (11) \end{aligned}$$

On the other hand, the histogram $H_\beta(E)$ of CV error $E(\boldsymbol{s})$ obtained by the EMC method with the inverse temperature β can be expressed by the expectation value over the probability distribution $p(\boldsymbol{s};\beta)$ as follows,

$$\begin{aligned} H_\beta(E) &= \sum_{\boldsymbol{s}} \delta\left(E - E(\boldsymbol{s})\right) p(\boldsymbol{s};\beta) && (12)\\ &= \sum_{\boldsymbol{s}} \delta\left(E - E(\boldsymbol{s})\right) \frac{\exp(-\beta E(\boldsymbol{s}))}{Z_\beta} && (13)\\ &= \frac{g(E)\exp(-\beta E)}{Z_\beta}. && (14) \end{aligned}$$

Consequently, the density of states $g(E)$ is given by using the normalization constant Z_β and the histogram $H_\beta(E)$,

$$g(E) = \frac{H_\beta(E)}{\exp(-\beta E)/Z_\beta}. \tag{15}$$

By using Eq. (11) and Eq. (15), when given the histogram $H_\beta(E)$ by the EMC method, we can calculate the density of states $g(E)$ by the iteration equation of the normalization constant Z_β and the density of states $g(E)$. In fact, since we can get the histograms $\{H_{\beta_m}(E)\}_{m=1}^{M}$ with several inverse temperatures $\{\beta_m\}_{m=1}^{M}$

by the EMC method, we can estimate the density of states $g(E)$ by the following iteration equations,

$$g(E) = \frac{\sum_{m=1}^{M} w_m^{-1} H_{\beta_m}(E)}{\sum_{m=1}^{M} w_m^{-1} n_m \exp(-\beta_m E)/Z_{\beta_m}}, \tag{16}$$

$$Z_{\beta_m} = \sum_{E} g(E) \exp(-\beta_m E), \tag{17}$$

where n_m is the total number of samples at β_m, and w_m is a weight factor originally determined by an autocorrelation time at β_m. The multiple histogram method defined by Eq. (16) is the best estimation method for minimizing the error in the resultant estimate for $g(E)$ [4]. In EMC method, however, the time correlation is difficult to define well, because the temperature of each replica does not remain constant during the simulation. Hence, we assume that the factor w_m is independent of temperature, i.e., $w_m = 1.0$ [9].

In principle, we can estimate the density of states $g(E)$ from a histogram $H_\beta(E)$ with a certain inverse temperature β. However, this leads to the poor accuracy of the density of the states $g(E)$. When we use a histogram $H_\beta(E)$ with a small value of inverse temperature β, we can search the wide range of energy, while the accuracy of the density of the states $g(E)$ goes worse in a small CV error. On the contrary, when we use a histogram $H_\beta(E)$ with a large value of inverse temperature β, it is difficult to estimate the density of states $g(E)$ in a whole range of CV error E because the algorithm searches the state with a small CV error. Consequently, in order to estimate the density of states $g(E)$ accurately, we need multiple histograms $\{H_{\beta_m}(E)\}_{m=1}^{M}$ with several values of inverse temperatures $\{\beta_m\}_{m=1}^{M}$. Therefore, the EMC method is good at estimating the density of states $g(E)$ by the multiple histogram method.

3. Simulation

In this section, we describe the simulation result in order to show the effectiveness of the proposed method.

3.1 Data and Issue with the Conventional Method

The data represented the firing rates of 23 neurons in the anterior inferior temporal (AIT) cortex of a monkey measured by conducting a single-unit recording, when the monkey was performing a sequential delayed matching-to-sample task requiring the identification of facial images [2], [3]. The presented images consisted of the face images of four different identities viewed from seven different angles. The AIT is known to be important for the face identification. Hence, the goal of this data analysis is to select the relevant neurons for the face identification from the neural recording data.

In this simulation, we treated the binary classification problem for face identification, Identity 1 vs. 3 pair and Identity 1 vs. 4 pair. The left figures in **Fig. 2** show the neural activity data for Identity 1 vs. 3 and Identity 1 vs. 4. The horizontal axes show the index of facial images, and the vertical ones the index of neuron. The center and the right figures respectively show the results for the logistic regression with the ARD prior [15], and those for the logistic regression with the L1 prior [13] for the data represented in the left figures. These two methods are well known as the sparse estimation for feature selection problem. We carry out LOOCV in each method. These figures show the selected features for each CV. The horizontal axis shows the index of CV, and the vertical one the index of neuron. The black cells in these figures indicate that the corresponding neurons were selected in each method, and the white cells indicate that the neurons were not selected.

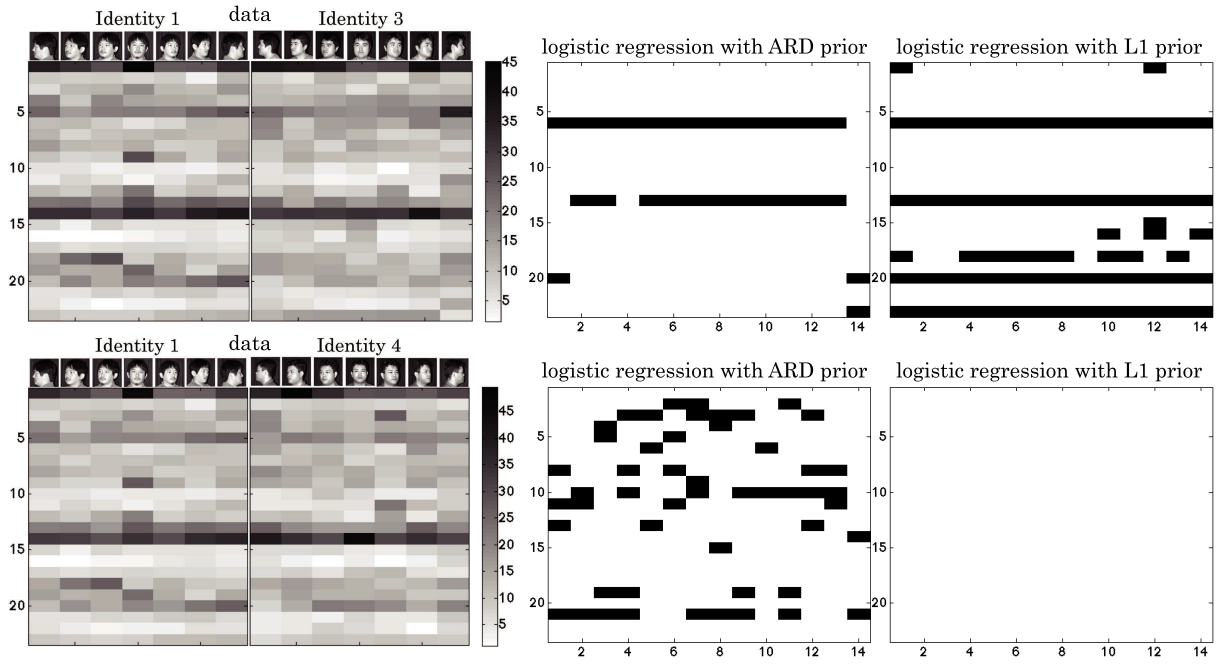

Fig. 2 The left, center, and right figures respectively show the neural activity data, the estimation results for the logistic regression with the ARD prior [15], and those for the logistic regression with the L1 prior [13]. The upper figures show the case of Identity 1 vs. 3, and the lower ones the case of Identity 1 vs. 4.

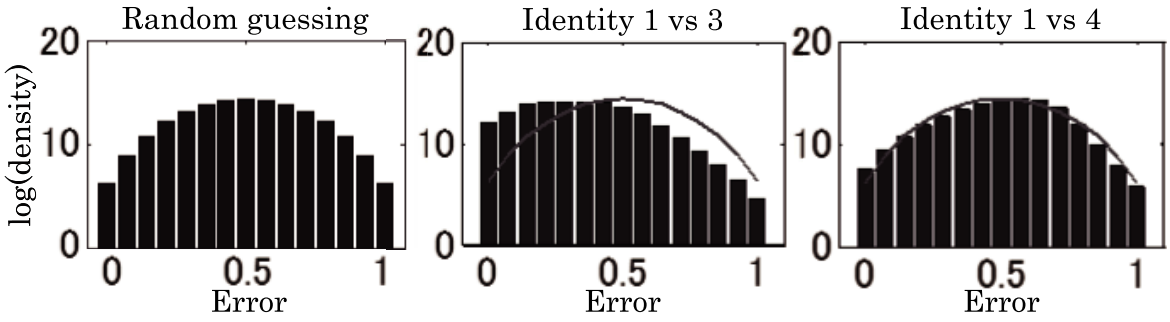

Fig. 3 The histograms of the CV error. The left, center, and right figure respectively show the histogram $g_{bi}(E)$ for random guessing, the histogram for Identity 1 vs. 3, and the histogram for Identity 1 vs. 4. The horizontal axes show the predictive error rate, and the vertical ones the logarithm of density. The curved lines in center and right figure show the histogram $g_{bi}(E)$ for random guessing.

From the result for Identity 1 vs. 3, though there are a little difference between the result of ARD prior and that of L1 prior, the 6th and 13th neurons are selected in both estimation methods. This implies the success for extracting the information by the feature selection. On the other hand, the results of the ARD prior for Identity 1 vs. 4 vary as the CV trial, and unstable. Moreover, the results of the L1 prior show that no neuron should be selected. Consequently, it is difficult to judge which neurons should be selected as the relevant neurons from the conventional estimation methods.

3.2 Verification for Comparing the Histograms

As the reason why the results in the above subsection were obtained, one can be considered that the given data do not include the information for binary classification. Although AIT is known to be important for face identification, it is difficult for us to judge whether the monkey used in the experiment recognizes the face identification between Identity 1 and Identity 4 from the neural data given. In this simulation, we calculated the histogram $g(E)$ by the CV for all combination of neurons, whose number was $2^{23} - 1 = 8{,}388{,}607$. **Figure 3** shows the results. The left, center, and right figure respectively show the histogram $g_{bi}(E)$ for random guessing, the histogram for Identity 1 vs. 3, and the histogram for Identity 1 vs. 4. The horizontal axes show the predictive error rate, and the vertical ones the logarithm of density. The curved lines in center and right figure show the histogram $g_{bi}(E)$ for random guessing. We also calculated the Kullback divergence $KL(g|g_{bi})$ for Identity 1 vs.3 and 1 vs. 4. As the result, the values of the Kullback divergence $KL(g|g_{bi})$ were 0.8104 for 1 vs. 3, and 0.0546 for 1 vs. 4.

From these results, the histogram $g(E)$ for Identity 1 vs. 3 is significantly different from the histogram $g_{bi}(E)$ for random guessing, in particular, in the small value of predictive error rate. This means that the neural activity data for Identity 1 vs. 3 include significant information for binary classification. On the other hand, the histogram $g(E)$ for Identity 1 vs. 4 cannot be seen the difference from the histogram $g_{bi}(E)$. This means that the neural activity data for Identity 1 vs. 4 do not include information for binary classification.

This conclusion is consistent to the result for sparse estimation above mentioned. That is, because the given data for Identity 1 vs. 4 do not include information for binary classification, the results of the sparse estimation for the case of Identity 1 vs. 4 vary or show that no neuron should be selected. Consequently, our proposed method gives us interpretation about the incomprehensible results of the conventional sparse estimation by judging whether the given data include information or not by exhaustively searching all combinations of features. This means the importance for clarifying the solution-space structure by the exhaustive search, and the effectiveness of our proposed method.

3.3 Estimation of Density of the States

Next, we show the simulation result for the estimation of the density of the states by using the EMC method and the Multiple histogram method.

We set the number M of replica as $M = 36$ for the EMC method, and the parameters $\{\beta_m\}_{m=1}^{M}$ of inverse temperature was set by a geometrical progression as follows,

$$\beta_m = r^{m-1} - 1, \tag{18}$$

$$r = (1 + 30.0)^{1/M}. \tag{19}$$

In this setting, $\beta_1 = 0.0$ and $\beta_M = 30.0$. The initial state of the combination $\boldsymbol{s}_m$ for each replica is randomly generate from the uniform distribution of $2^{23} - 1 = 8388607$ combinations. We define the Monte Carlo step (MCS) as carrying out the step 1 and step 2 of the EMC method once, and we simulate 2,300 MCS. Then, we calculate the CV error $36 \times 2{,}300 = 82{,}800$ times. Its computational cost is about $\frac{1}{100}$ of the exhaustive search method, because the number of all combination of features is 8,388,607. The initial condition of the iterative equations, Eqs. (16) and (17), for the multiple histogram method is set as $Z_{\beta_m} = e^{-1}$ for all β_m, which corresponds to the setting of free energy $F_{\beta_m} \equiv -\log Z_{\beta_m}$ as $F_{\beta_m} = 1$.

Figure 4 (a) shows the result of the estimation of the density of the states $g(E)$. The horizontal axis shows the logarithm of the density and the vertical one shows the value of CV error. Open circles in Fig. 4 (a) show the density of the states $g(E)$ calculated by the exhaustive search method, and crossing points show the estimated density of the states by using the EMC method and the Multiple histogram method. In Fig. 4 (a), the typical number of iteration to converge $g(E)$ is 50 steps. From this result, the estimated density of the states is almost same as the density of the states $g(E)$ calculated by the exhaustive search method. This shows the effectiveness of the proposed estimation method of the

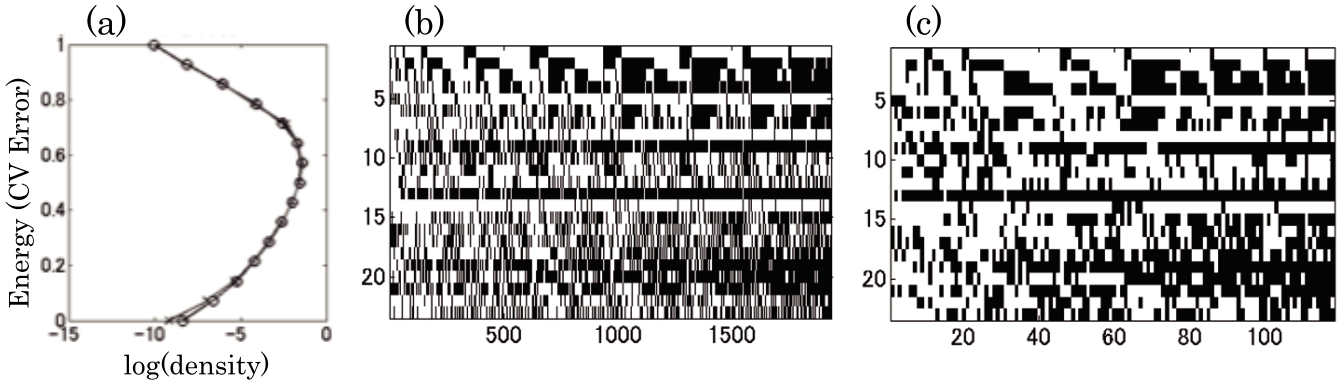

Fig. 4 Estimated histogram of CV error in Fig(a). Number of CV error calculation is 82,800. Open circle is the true value obtained by the exhaustive search with all combinations, in which the number of CV error calculation is 8,388,607, and the crossing point is the estimated value. Feature subsets having $E(\boldsymbol{s}) = 0$ obtained by the simulation of (b) exhaustive search and (c) the EMC simulation. Horizontal and vertical axes represent number of subsets and neurons. Black cells indicate that the feature is in the subset.

density of the states $g(E)$.

Figure 4 (b) and 4 (c) respectively show the combination of the selected features $\boldsymbol{s}$ with minimum CV error $E(\boldsymbol{s})$, i.e., $E(\boldsymbol{s}) = 0$, obtained by the exhaustive search method and obtained by the EMC method. As a result of the exhaustive search method, the number of combinations of features with $E(\boldsymbol{s}) = 0$ was 1,938. On the contrary, the EMC method can search the 118 combinations of features with $E(\boldsymbol{s}) = 0$, which is 6.09% of all combinations. This indicates the goodness of the EMC method because the calculation cost of the EMC method was about 1% of the exhaustive search method. We can also see a similar tendency for selected features, i.e., neurons between the exhaustive search method and the EMC method. For example, from Fig. 4 (b), the 9th and 13th neurons are selected in many combinations, and the 5th, 8th, and 14th neurons are hardly selected. This tendency about the selected neurons can be seen from the result for the EMC method.

4. Discussion and Conclusion

In this paper, we focused on the feature selection problem for the binary classification problem with linear discriminant. We proposed a method to judge whether the given data include information for binary classification problem by calculating the histogram of the generalization error for all combination of features by the cross validation method. We applied the proposed method to the feature selection problem for selecting the relevant neurons for face identification. As a result, we clarify that the data for a certain pair of identity 1 vs. 4 does not include information for binary classification. This result gives us a clear reason why the solution in a conventional sparse estimations such as LASSO and ARD become unstable, and indicates the importance for exhaustively searching all combination of features. We also proposed the method to efficiently estimate the density of the states by using the exchange Monte Carlo method and the multiple histogram method, and showed its effectiveness by comparing the result of the proposed estimation method to that of the exhaustive search method.

The reason why we obtained the result that the data for the pair of identity 1 vs. 4 does not include information for binary classification can be the small size of samples. As above mentioned in Section 1, Cover and Van Campenhout showed that showed that no non-exhaustive sequential feature selection procedure can be guaranteed to produce the optimal subset in the feature selection problem. Moreover, they also showed that the exhaustive search method has a risk to provide the subset of features different from the optimal subset when the sample size is small [1]. However, the result of our simulation showed that the data for the pair of identity 1 vs. 3 includes information for binary classification although the sample size is same as that for the pair of identity 1 vs. 4. Therefore, the method to judge whether the given data set have information for feature selection is very important, and the result of our study indicates that our proposed method is efficient.

As the future work, it is important whether the relevant features are selected from the result of the exhaustive search method. One can consider the method to select the relevance features which are selected in many combinations for the exhaustive search method, as described above in Fig. 4 (b) and in Fig. 4 (c). However, it is the problem which combinations are selected as the relevant combination of features. In the results in Fig. 4 (b) and in Fig. 4 (c), we focused on the combination of neurons with minimum CV error, i.e., $E(\boldsymbol{s}) = 0$. However, this selection has arbitrariness. Hence, in order to select the combination of features without arbitrariness as much as possible, we need a test for statistical significance for each combination of features. This test have to carry out many times, and we have to consider the correction by the multiple hypothesis testing [5]. Consequently, Development of the feature selection method by using the exhaustive search method and the multiple hypothesis testing is very important, and should be addressed as the future works.

References

[1] Cover, T.M. and Van Campenhout, J.M.: On the Possible Orderings in the Measurement Selection Problem, *IEEE Trans. Systems, Man, and Cybernetics*, Vol.7, pp.657–661 (1977).
[2] Eifuku, S., De Souza, W.C., Tamura, R., Nishijo, H. and Ono, T.: Neuronal Correlates of Face Identification in the Monkey Anterior Temporal Cortical Areas, *J Neurophysiol.*, Vol.91, pp.358–371 (2004).
[3] Eifuku, S., De Souza, W.C., Nakata, R., Ono, T. and Tamura, R.: Neural Representations of Personally Familiar and Unfamiliar Faces in the Anterior Inferior Temporal Cortex of Monkeys, *PLoS One*, Vol.6,

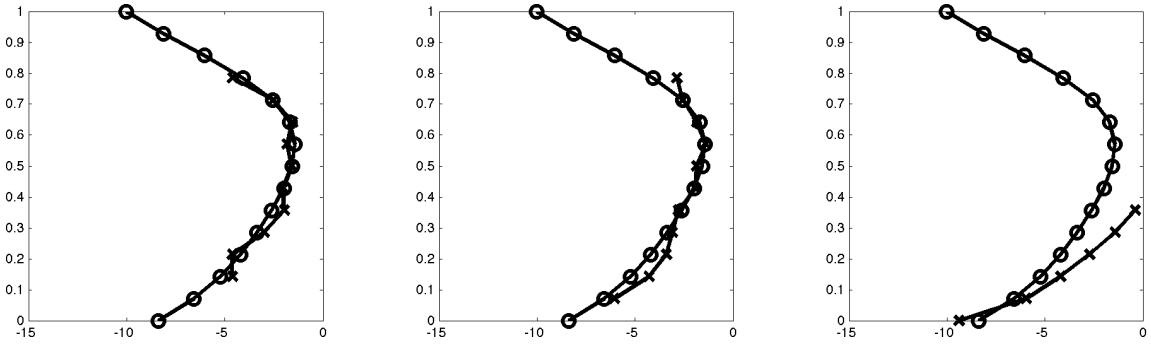

Fig. A·1 Estimation result of the histogram of CV error for the single temperature. The values of temperature β are from left to right $\beta = 0.0$, $\beta = 5.4505$ and$\beta = 30.0$. In each figure, the circles indicate the histogram obtained by the exhaustive search method, and the crossing points indicate the estimated histogram.

e18913 (2011).

[4] Ferrenberg, A.M. and Swandsen, R.H.: Optimized Monte Carlo Data Analysis, *Phys. Rev. Lett.*, Vol.63, pp.1195–1198 (1989).

[5] Goeman, J.J. and Solari, A.: Multiple hypothesis testing in genomics, *Statistics in Medicine*, Vol.2014, No.33, pp.1946–1978 (2014).

[6] Guyon, I. and Elisseeff, A.: An Introduction to Variable and Feature Selection, *J. Mach. Learn. Res.*, Vol.3, pp.1157–1182 (2003).

[7] Haury, A.-C., Gestraud, P. and Vert, J.-P.: The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures, *PLoS One*, Vol.6, No.12, e28210 (2011).

[8] Hukushima, K. and Nemoto, K.: Exchange Monte Carlo Method and Application to Spin Glass Simulations, *J. Phys. Soc. Jpn.*, Vol.65, pp.1604–1608 (1996).

[9] Hukushima, K.: Extended ensemble Monte Carlo approach to hardly relaxing problems, *Comput. Phys. Commun*, Vol.147, pp.77–82 (2002).

[10] Jain, A.K., Duin, R.P.W. and Mao, J.: Statistical Pattern Recognition: A Review, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.22, pp.4–37 (2000).

[11] Kalousis, A., Prados, J. and Hilario, M.: Stability of feature selection algorithms: a study on high-dimensional spaces, *Knowl. Inf. Syst.*, Vol.12, No.11, pp.95–116 (2007).

[12] Tibshirani, R.: Regression shrinkage and selection via the lasso, *J. Royal Stat. Soc. B*, Vol.58, No.1, pp.267–288 (1996).

[13] Tomioka, R. and Sugiyama, M.: Dual Augmented Lagrangian Method for Efficient Sparse Reconstruction, *IEEE Signal Processing Letters*, Vol.16, No.12, pp.1067–1070 (2009).

[14] Xu, H. and Mannor, S.: Sparse Algorithms are Not Stable: A No-Free-Lunch Theorem, *IEEE Trans. Pattern Anal. Mach, Intell.*, Vol.34, No.1, pp.187–193 (2012).

[15] Yamashita, O., Sato, M.A., Yoshioka, T., Tong, F. and Kamitani, Y.: Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns, *Neuroimage*, Vol.42, No.4, pp.1414–1429 (2008).

Appendix

A.1 Support Vector Machine

In Appendix A, we describe the learning of support vector machine (SVM), which is used as the linear discriminant in this study. The SVM learns the parameter $\boldsymbol{w}$ and b based on the principle of maximization of margin, which is the distance between the decision boundary and the closest sample to the decision boundary. More concretely, the learning of SVM is formulated by using the Lagrange multiplier,

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \max_{\boldsymbol{\lambda},\boldsymbol{\mu}} \left\{ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\lambda_i\left\{t_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) - 1\right\} - \sum_{i=1}^{N}\mu_i\xi_i \right\}, \qquad \text{(A.1)}$$

where $\boldsymbol{\xi} = \{\xi_i\}_{i=1}^{N}, \xi_i \geq 0$ is slack variable which represents a penalty for misclassification, and the variable $\boldsymbol{\lambda}, \boldsymbol{\mu}$ is the Lagrange multiplier. The variable C is a regularization constant which controls the trade-off between the penalty of slack variable $\boldsymbol{\xi}$ and the margin maximization. In our simulation, we set $C = 5.0$.

A.2 Verification of Multiple Histogram Method

In Appendix B, we show the result of the multiple histogram method for single temperature in order to verify the effectiveness of multiple histogram method. **Figure A·1** show the estimation result of the histogram of CV error for the single temperature. The values of temperature β are from left to right $\beta = 0.0$, $\beta = 5.4505$ and $\beta = 30.0$. These figures indicate that the estimation for the single temperature partially approximates the histogram of CV error. This is caused that the Monte Carlo algorithm is difficult to sample from whole region of CV error in a single temperature. As can be seen from Eq. (15), if $H_\beta(E) = 0$, then $g(E) = 0$. Multiple histogram method overcomes the problem that it estimates $g(E) = 0$ by using the multiple sequence of Markov chain with different temperatures.

Kenji Nagata is an assistant professor in Graduate School of Frontier Sciences, The University of Tokyo, Japan. He received his master and doctoral degrees on computer science from Tokyo Institute of Technology in 2006 and in 2008, respectively. His research interest is in theory and applications of machine learning, in particular, Bayesian learning theory, Markov chain Monte Carlo method, and its architecture.

Jun Kitazono received his B.L.A., M.S. and Ph.D. degrees from The University of Tokyo in 2008, 2010 and 2013, respectively. He was a predoctoral research fellow of JSPS from 2010 to 2013, a postdoctoral researcher at "Okanoya Emotional Information Project," ERATO, Japan Science and Technology Agency (JST) from 2013 to 2014, and a postdoctoral researcher at The University of Tokyo in 2014. He is currently an assistant professor at the Graduate School of Engineering, Kobe University. His research interests include machine learning and computational neuroscience.

Shinichi Nakajima is a senior researcher in Berlin Big Data Center, Machine Learning Group, Technische Universität Berlin. He received his master degree on physics in 1995 from Kobe University, and worked with Nikon Corporation until September 2014 on statistical analysis, image processing, and machine learning. He received his doctoral degree on computer science in 2006 from Tokyo Institute of Technology. His research interest is in theory and applications of machine learning, in particular, Bayesian learning theory, computer vision, and quantum chemistry.

Satoshi Eifuku is a professor and chairman of the Department of System Neuroscience, Faculty of Medicine, Fukushima Medical University, Japan. He received his doctoral degree on medical science in 1994 from Toyama Medical and Pharmaceutical University. His current research field is in general neuroscience and nerve/muscular physiology. His research interests include neural basis for social cognition and behavior and neural basis for visual recognition and memory.

Ryoi Tamura is a professor in Department of Physiology, University of Toyama, Japan. He received his doctoral degree on medicine in 1996 from University of Toyama. His research field is neuroscience and neurophysiology.

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