frontiers in Neuroinformatics

METHODS
published: 15 August 2018
doi: 10.3389/fninf.2018.00051

Sparse Ordinal Logistic Regression and Its Application to Brain Decoding

Emi Satake[1†], *Kei Majima*[1†], *Shuntaro C. Aoki*[2] *and Yukiyasu Kamitani*[1,2*]

[1] *Graduate School of Informatics, Kyoto University, Kyoto, Japan,* [2] *ATR Computational Neuroscience Laboratories, Kyoto, Japan*

Brain decoding with multivariate classification and regression has provided a powerful framework for characterizing information encoded in population neural activity. Classification and regression models are respectively used to predict discrete and continuous variables of interest. However, cognitive and behavioral parameters that we wish to decode are often ordinal variables whose values are discrete but ordered, such as subjective ratings. To date, there is no established method of predicting ordinal variables in brain decoding. In this study, we present a new algorithm, sparse ordinal logistic regression (SOLR), that combines ordinal logistic regression with Bayesian sparse weight estimation. We found that, in both simulation and analyses using real functional magnetic resonance imaging (fMRI) data, SOLR outperformed ordinal logistic regression with non-sparse regularization, indicating that sparseness leads to better decoding performance. SOLR also outperformed classification and linear regression models with the same type of sparseness, indicating the advantage of the modeling tailored to ordinal outputs. Our results suggest that SOLR provides a principled and effective method of decoding ordinal variables.

Keywords: decoding, functional magnetic resonance imaging, ordinal logistic regression, bayesian estimation, sparseness

OPEN ACCESS

Edited by:
Pedro Antonio Valdes-Sosa, Clinical Hospital of Chengdu Brain Science Institute, China

Reviewed by:
Tonio Ball, Translational Neurotechnologie Labor, Albert-Ludwigs-Universität Freiburg, Germany
Felix Carbonell, Biospective Inc., Canada
Jorge Bosch-Bayard, Instituto de Neurobiología, Universidad Nacional Autonoma de Mexico, Mexico

***Correspondence:**
Yukiyasu Kamitani
kamitani@i.kyoto-u.ac.jp

[†]These authors have contributed equally to this work

Received: *22 December 2017*
Accepted: *24 July 2018*
Published: *15 August 2018*

Citation:
Satake E, Majima K, Aoki SC and Kamitani Y (2018) Sparse Ordinal Logistic Regression and Its Application to Brain Decoding. Front. Neuroinform. 12:51. doi: 10.3389/fninf.2018.00051

INTRODUCTION

Application of multivariate classification and regression models to functional magnetic resonance imaging (fMRI) signals has allowed the extraction of information encoded in population neural activity. Classification models are used to predict categorical variables, such as discrete stimuli and task conditions (Haynes and Rees, 2006; Norman et al., 2006; Pereira et al., 2009), while regression models are used to predict continuous parameters of interest (Cohen et al., 2011). These two types of prediction model are employed depending on the type of variable we wish to decode.

However, variables we attempt to decode are often ordinal—discrete variables whose values (classes) are ordered. For example, behavioral ratings that quantify subjective states such as the emotional feeling, impression, and preference (e.g., Chu et al., 2011; Valente et al., 2011; Baucom et al., 2012; Smith et al., 2014; Chang et al., 2015) are discrete and ordered. The intervals between classes are not defined in many cases. Furthermore, parameters of stimuli used in experiments have often been restricted to take discrete values (e.g., Kamitani and Tong, 2005, 2006; Miyawaki et al., 2008; Staeren et al., 2009; Nishio et al., 2012). Even when a parameter is defined in a metric space, the distributions of the voxel patterns are not necessarily proportionally spaced between the discrete values. Thus, discretized parameters could be better treated as ordinal variables.

Such variables have been predicted with classification and regression models in previous decoding studies. In studies using classification models, the given discrete levels were treated as nominal classes and classification models were trained to classify input brain activity patterns into one of those classes (e.g., Miyawaki et al., 2008). In studies using regression models, models were trained by treating a given ordinal variable as a continuous variable, and continuous outputs from the models were then used as the prediction results (e.g., Chu et al., 2011; Valente et al., 2011; Nishio et al., 2012; Chang et al., 2015).

Owing to the nature of ordinal variables, classification and regression models are not considered appropriate for ordinal variable prediction. An ordinal variable is a discrete variable whose classes are ordered. By definition, the distances between different classes are not given, and only the relative ordering between classes is important. In handling rating scores, for example, level 2 is placed between level 1 and level 3, but the magnitudes of the differences between levels are undefined. When a regression model is fitted using class numbers as labels, the distances between consecutive classes are treated as equal. Hence, the resultant fitness of the model depends on those deceptive distances. Meanwhile, classification models assume classes to be nominal categories and ignore given relative similarities between classes that provide helpful information for constructing a model with better prediction performance.

We here present an approach using ordinal regression, a type of generalized linear modeling whose output variable is assumed to be ordinal (Winship and Mare, 1984). In ordinal regression, similar to linear regression, a linear combination of input variables is used to predict the target variable (**Figure 1A**). Differently from linear regression, however, the value of the linear combination for a given input sample is not directly used as the prediction. In ordinal regression, a set of thresholds is introduced to divide the real number line into disjoint segments. These segments correspond to the discrete classes of the target variable. The class corresponding to the segment where the value of the linear combination lies is then selected as the prediction. This treats the class number as a discrete variable without using the metric in the space of the output variable. By tuning both linear weights and thresholds, ordinal regression models can be better fitted to given ordinal data than linear regression models.

Compared with classification models, ordinal regression models are expected to be efficient in learning, leading to better prediction performance. Classification models learn decision boundaries that are used to classify input samples into classes in the feature space (**Figure 1B**, left). Their degree of freedom increases as the number of classes increases, which makes parameter estimation sensitive to noise. In contrast, all decision boundaries of an ordinal regression model are restricted to be orthogonal to a single line in the feature space, and the degree of freedom is smaller than that for classification models (**Figure 1B**, right). This lower complexity of ordinal regression models reduces the chance of overfitting and leads to better generalization performance than classification models.

To introduce a multivariate prediction model into fMRI decoding analysis, it is generally important to choose an appropriate set of input voxels because the presence of many irrelevant voxels can lead to poor generalization performance due to overfitting. In standard decoding analysis, only tens or hundreds of fMRI samples are obtained to train the prediction model, while the input feature vector consists of thousands of voxels. Thus, overfitting readily occurs if all available voxels are used as input features. To solve this problem, our previous study proposed a classification algorithm that simultaneously performs voxel selection and parameter estimation, and demonstrated that the method successfully prevents overfitting in the presence of many irrelevant voxels (Yamashita et al., 2008). In that study, a Bayesian extension of logistic regression was proposed where the automatic relevance determination (ARD; MacKay, 1992; Neal, 1996) prior was used as the prior distribution of the weight vector. This resulted in selecting a small number of voxels as important by estimating the corresponding weight parameters to be nonzero, and ignoring the other voxels by estimating their weight parameters to be zero. This sparse parameter estimation provided a method of voxel selection by virtually eliminating voxels associated with zero-valued weight parameters.

In the present study, we combine ordinal regression with the sparse estimation (**Figure 1C**) to build an ordinal prediction model suited to fMRI decoding. As our model is based on ordinal logistic regression (OLR; McCullagh, 1980), a standard ordinal regression model, we refer to our proposed method as sparse ordinal logistic regression (SOLR). We evaluate the performance of SOLR using both simulation and real fMRI data. In these analyses, the prediction performance of SOLR is compared with that of an OLR model without a sparseness constraint to examine the utility of the sparseness. Likewise, the prediction performance is compared with that of regression and classification models having the same type of sparseness, sparse linear regression (SLiR; Tipping, 2001; Bishop, 2006) and sparse multinomial logistic regression (SMLR; Yamashita et al., 2008), to examine the superiority of SOLR in ordinal variable prediction. To examine whether SOLR works well in practical situations, we compare the decoding performances for different numbers of training samples and input dimensions in the simulation analysis. In the analysis using real fMRI data, we tested the four previously mentioned algorithms on a dataset taken from Miyawaki et al. (2008). In this previous study, stimulus images were reconstructed from fMRI responses by training decoders on 440 samples using about 1,000 voxels from the primary visual cortex (V1) in both hemispheres as input. Using this dataset, we demonstrate that SOLR better predicts ordinal variables in a practical situation of fMRI decoding analysis.

MATERIALS AND METHODS

Algorithm

This section first describes OLR, which is a generalized linear model for ordinal dependent variables (McCullagh, 1980; Winship and Mare, 1984), and then explains SOLR by introducing a Bayesian framework to estimate parameters. OLR with L2-regularization (L2OLR) is also explained in this section.

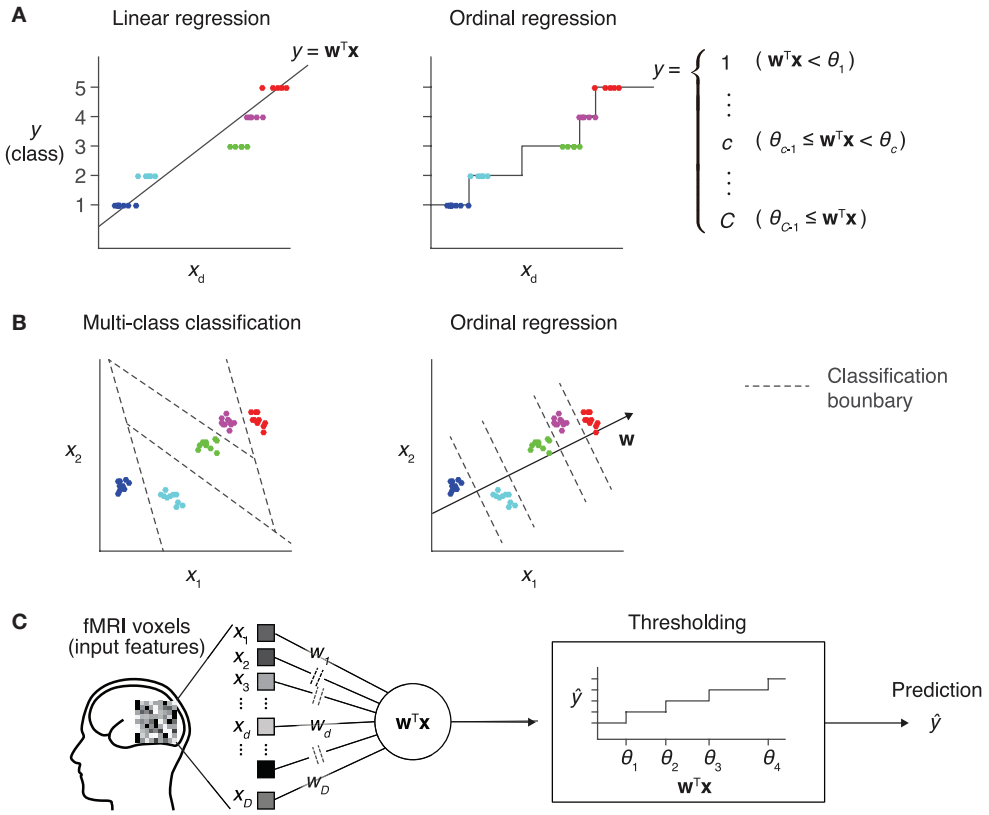

FIGURE 1 | Prediction models for an ordinal dependent variable. **(A)** Comparison between linear regression and ordinal regression. In linear regression, the prediction is calculated as a linear combination of input features. Here, y is assumed to be an ordinal variable and it takes one of five levels. The regression line on a single input dimension (x_d) is shown (left). In ordinal regression, a linear combination is used for prediction, but not directly treated as its prediction. Thresholds that divide the real number line into five disjoint segments are introduced as parameters. Each segment corresponds to one of the five levels, and the prediction is obtained as the level corresponding to the segment that the linear combination value belongs to (right). **(B)** Comparison between multiclass classification and ordinal regression. In prediction with a multiclass classification model, classification boundaries that separate different classes are learned in the feature space (left). In ordinal regression, classification boundaries are restricted to be orthogonal to one single vector, which reduces the degree of freedom compared with typical classification models (right). **(C)** Sparse ordinal logistic regression (SOLR). A standard ordinal regression model, the ordinal logistic regression model, is combined with a Bayesian sparse parameter estimation method, and is introduced into fMRI decoding analysis in the present study. By estimating a subset of weight parameters to be nonzero and the other weight parameters to be zero, this method simultaneously selects important input features (voxels) and estimates the weight parameters. In prediction, it linearly combines voxel values with the estimated sparse weight vector, then outputs one of the given levels by comparing the resultant value with thresholds.

Our MATLAB and Python implementations of SOLR and L2OLR are available at our Github respository[1]

OLR is one of the generalized linear models whose dependent variable (or target variable to be predicted) is assumed to be an ordinal variable. In OLR, the dependent variable $y \in \{1, \cdots, C\}$ is assumed to follow the underlying process given by

$$z = \mathbf{w}^{\mathrm{T}}\mathbf{x} + \varepsilon, \tag{1}$$

$$y = \begin{cases} 1 & (z < \mu_1) \\ \vdots & \vdots \\ c & (\mu_{c-1} \leq z < \mu_c), \\ \vdots & \vdots \\ C & (\mu_{c-1} \leq z) \end{cases} \tag{2}$$

where $\mathbf{x} = (x_1, \cdots, x_D)^{\mathrm{T}} \in \mathbb{R}^D$ is the vector of D independent variables (input features, or voxel values), is the linear weight vector, ε is a random variable representing the noise, z is a latent variable assumed to link the dependent and independent variables in the model, and $\mu_1, \mu_2, \cdots, \mu_{C-1}$ $(\mu_1 \leq \mu_2 \leq \cdots \leq \mu_{C-1})$ are threshold parameters. In OLR, ε is assumed to follow the logistic distribution with a mean of zero and a variance of 1. The threshold parameters are collectively denoted by a single vector $\boldsymbol{\mu}$.

In OLR, $\mathbf{w}$ and $\boldsymbol{\mu}$ are estimated by maximizing the log likelihood function with a gradient method. The log likelihood with respect to $\mathbf{w}$ and $\boldsymbol{\mu}$ is given by

$$\log p\left(\mathbf{Y}|\mathbf{w}, \boldsymbol{\mu}, \mathbf{X}\right) = \sum_{n=1}^{N}\sum_{c=1}^{C} y_{nc} \log\left[F\left(\mu_c - \mathbf{w}^{\mathrm{T}}\mathbf{x}_n\right) - F\left(\mu_{c-1} - \mathbf{w}^{\mathrm{T}}\mathbf{x}_n\right)\right], \tag{3}$$

[1]https://github.com/KamitaniLab/SOLR

where, y_{nc} is a binary variable indicating whether the value of the dependent variable for the n-th sample is c. y_{nc} is set to 1 if the value of the dependent variable for the n-th sample is c and y_{nc} is set to zero otherwise (1-of-k representation). $\mathbf{x}_n$ is the vector of feature values for the n-th sample. N samples are used for estimation, and are collectively denoted by the $N \times C$ matrix $\mathbf{Y}$ and the $N \times D$ matrix $\mathbf{X}$. The function F is the logistic sigmoid function defined by

$$F(z) = \frac{1}{\left(1 + e^{-z}\right)}. \tag{4}$$

In the above log likelihood function, the first and last threshold parameters μ_0 and μ_C are respectively set to $-\infty$ and $+\infty$ by convention.

We next introduce a Bayesian framework to estimate the above parameters sparsely. We introduce prior distributions for parameters to be estimated in OLR. For the parameter $\mathbf{w}$, we assume that

$$\mathbf{w}|\boldsymbol{\alpha} \sim \mathcal{N}\left(\mathbf{0}, diag\left({\alpha_1}^{-1}, \cdots, {\alpha_D}^{-1}\right)\right), \tag{5}$$

where $\alpha_1, \cdots, \alpha_D$ are hyperparameters that determine the importance of voxels and are called relevance parameters. They are collectively denoted by the single vector $\boldsymbol{\alpha}$. $\mathcal{N}(\mathbf{m}, \Sigma)$ represents the multidimensional Gaussian distribution with mean $\mathbf{m}$ and covariance Σ. $\mathbf{0}$ represents the zero vector, while $diag\left({\alpha_1}^{-1}, \cdots, {\alpha_D}^{-1}\right)$ represents the diagonal matrix whose diagonal elements are ${\alpha_1}^{-1}, \cdots, {\alpha_D}^{-1}$ and non-diagonal elements are zero. If ${\alpha_d}^{-1}$ is small, the distribution function of w_d has a sharp peak around zero and the corresponding voxel thus tends to be virtually ignored in prediction. If ${\alpha_d}^{-1}$ is large, w_d can take a large value. For α_d, we further assume the non-informative prior whose distribution is given by

$$p(\alpha_d) = {\alpha_d}^{-1}, \tag{6}$$

as often adopted in previous studies (Yamashita et al., 2008). Additionally, for the parameter $\boldsymbol{\mu}$, we assume the non-informative prior, which is expressed as

$$p(\boldsymbol{\mu}) = \lim_{\sigma^2 \to \infty} \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}\right). \tag{7}$$

We then obtain the log of the posterior distribution as

$$\begin{aligned} &\log p(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\alpha}|\mathbf{X}, \mathbf{Y}) \\ &\propto \log p(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\alpha}, \mathbf{Y}|\mathbf{X}) \\ &= \log p(\mathbf{Y}|\mathbf{w}, \boldsymbol{\mu}, \mathbf{X}) + \log p(\mathbf{w}|\boldsymbol{\alpha}) + \log p(\boldsymbol{\alpha}) + \log p(\boldsymbol{\mu}). \end{aligned} \tag{8}$$

In the present study, the values of $\mathbf{w}$, $\boldsymbol{\mu}$, and $\boldsymbol{\alpha}$ that maximize the above function—the maximum a posteriori (MAP) solution—were estimated with training data, and the values of $\mathbf{w}$ and $\boldsymbol{\mu}$ were then used in prediction on test data. Because the MAP solution for the above cannot be derived in a closed form, we used the mean-field variational Bayesian approximation and the Laplace approximation (Attias, 1999, Bishop, 2006; see **Appendix**). Once we obtain the MAP solution, we can calculate the predictive probability of each class for a given new input vector. The class with the highest predictive probability was chosen as the prediction outcome.

To examine the effect of voxel selection by ARD, we compared the performance of SOLR with that of OLR having L2-regularization (L2OLR). In L2OLR, we assume that the prior distribution of $\mathbf{w}$ is the Gaussian distribution with zero mean and isotropic covariance, as expressed by

$$\mathbf{w}|\alpha \sim \mathcal{N}\left(\mathbf{0}, \alpha^{-1}\mathbf{I}\right). \tag{9}$$

α is a hyperparameter that controls the degree of regularization. In a similar manner to SOLR, we assume non-informative priors for α and $\boldsymbol{\mu}$. We estimated the MAP solution with the same approximation techniques, and then made a prediction using the estimated parameters.

Simulation Analysis

We compared the prediction performance across SOLR, L2OLR, SLiR, and SMLR using simulation data. For data generation, five D-dimensional Gaussian distributions were prepared, and samples for class c were generated from the c-th Gaussian distribution. The means of the Gaussian distributions were given by

$$\boldsymbol{\mu}_1 = (0, \cdots, 0)^{\mathrm{T}}, \tag{10}$$

$$\boldsymbol{\mu}_c = \boldsymbol{\mu}_{c-1} + \left(h_{c,1}, \cdots, h_{c,d}, \cdots, h_{c,10}, 0, \cdots, 0\right)^{\mathrm{T}}, \tag{11}$$

where $h_{c,d}$ ($c = 2, \cdots, 5$; $d = 1, \cdots, 10$) are parameters that specify the intervals between the means, and each was sampled from an exponential distribution with a mean of 1.0. Only the first 10 dimensions have information on classes, and the other dimensions are irrelevant. In each of the first 10 dimensions, the mean of the input feature monotonically increases against the class label, which leads to an ordinal structure in the feature space. We also conducted the same simulation analysis in the case that the means of the Gaussian distributions are equally spaced by setting all $h_{c,d}$ to 1.0, and observed qualitatively similar comparison results. The covariance matrices of the Gaussian distributions were set as diagonal matrices regardless of the class label. The standard deviation in each dimension was set to 3.0. D was set to 25, 50, 100, 250, 500, 1,000, 1,500, and 2,000 to characterize the prediction performance as a function of the number of input dimensions.

To evaluate the prediction performance, a prediction model was trained and tested on independent sets of samples. N samples and 1,000 samples were respectively generated from the same Gaussian distributions as training and test data. N was set to 10, 25, 50, 100, 250, 500, 1,000, 1,500, and 2,000 to characterize the prediction performance as a function of the number of training samples. Equal numbers of samples were generated for the five classes. To quantify the prediction performance, we calculated the Spearman rank correlation between true and predicted labels using test data. The same simulation procedure was repeated 100 times, and the prediction performance measured by the Spearman rank correlation was averaged across those 100 repetitions.

fMRI Data Analysis

We compared the prediction performance across the four algorithms using real fMRI data from Miyawaki et al. (2008). The dataset can be downloaded from public databases[2] (Poldrack et al., 2013; Takemiya et al., 2016). It contains fMRI signals when the subject was viewing visual images consisting of contrast-defined 10 × 10 checkerboard patches. Each patch was either a flickering checkerboard or a homogeneous gray area. The dataset consists of two independent sessions. One is a random image session, in which a spatially random pattern was presented for 6 s and there was a subsequent 6-s rest period. A total of 440 different random patterns were presented to the subject. The other is a figure image session, where a letter of the alphabet or a simple geometric shape was presented for 12 s and there was a subsequent 12-s rest period. Five letters of the alphabet and five geometric shapes were presented eight times.

Miyawaki et al. (2008) successfully reconstructed presented images from fMRI responses by combining multiple classifiers. To reconstruct arbitrary visual images, a set of local regions that cover the entire stimulus image area was predefined, the mean contrast in each local region was then predicted by a classifier, and the outputs from the classifiers for those local regions were then optimally combined to produce a single reconstructed image. In the previous study, SMLR was used to construct classifiers. Here, we used SOLR, L2OLR, SLiR, and SMLR for contrast prediction, and compared the prediction performance among them. In this analysis, the amplitudes of the 996 voxels in the primary visual cortex (V1) in both hemispheres were used as inputs. The V1 voxels were identified by the standard retinotopy mapping analysis (see Miyawaki et al., 2008).

RESULTS

Simulation Analysis

In the simulation analysis, samples from the five classes were generated from five multidimensional Gaussian distributions, and a prediction model using each algorithm was trained and tested on independent sets of data samples (see section Materials and Methods). We set only the first 10 dimensions to have information on the classes while keeping the other dimensions irrelevant. In each of the first 10 dimensions, the centers of the five Gaussian distributions were placed so that the mean of the input feature monotonically increases against the class number to assume an ordinal structure in the feature space.

To characterize prediction performance when the number of input dimensions is large, performance was calculated as a function of the number of input dimensions (**Figure 2A**). Prediction performance was evaluated using the Spearman rank correlation between true and predicted labels. The number of training samples was fixed to 100, which is a typical size in real fMRI decoding analysis. While all algorithms had similar performance when the number of input dimensions was small, SOLR outperformed the other algorithms as the number increased.

Furthermore, the prediction performance was calculated as a function of the number of training samples to characterize the performance when the number of training data is small (**Figure 2B**). Here, the number of input dimensions was fixed to 1,000, which is a typical input size in decoding analysis. As a result, SOLR had higher performance than the other algorithms when the number of training data was small. As we increased the number of training samples, all algorithms reached similar accuracies.

fMRI Data Analysis

We also evaluated the prediction performance using the real fMRI dataset from Miyawaki et al. (2008). The cited study measured fMRI responses as the subject viewed 10 × 10 binary images and successfully reconstructed arbitrary visual images from the fMRI responses (**Figures 3A, B**). In the reconstruction procedure, it was assumed that a stimulus image can be represented by a linear combination of local image bases of multiple scales (1 × 1, 1 × 2, 2 × 1, and 2 × 2; **Figure 3A**). There are a few possible stimulus states (contrast patterns) in the region specified by a single image basis, and the stimulus states can be classified according to the mean contrasts (**Figure 3C**). The mean contrast for the 1 × 1 image basis is binary, but the mean contrast for 1 × 2 and 2 × 1 image bases takes one out of three discrete values, and that for the 2 × 2 image basis takes one out of five discrete values. While the intervals between the successive mean contrasts can be regarded as equal in the contrast space, the distributions of voxel patterns for the discrete values are not necessarily equally spaced. In fact, the amplitude of the most responsive voxel in V1 for each image basis monotonically increases against but is not proportional to the mean contrast level (**Figure 3D**). Miyawaki et al. (2008) predicted the mean contrasts for the image bases using binary or multiclass classifiers based on sparse logistic regression (sparse [multinomial] logistic regression, SMLR), disregarding the order of contrasts. Here, we used SOLR, L2OLR, SLiR, and SMLR to predict the mean contrasts for 1 × 2, 2 × 1, and 2 × 2 image bases, and compared the prediction performance among them.

We evaluated the prediction performance of each algorithm by five-fold cross-validation using the random session data (**Figure 4A**). The prediction performance for each image basis was quantified by the Spearman rank correlation between true and predicted contrasts across test samples. The continuous outputs of SLiR were assigned to the nearest discrete labels. For all basis shapes, SOLR outperformed the other algorithms. The median performances of SOLR were significantly higher than those of L2OLR ($p < 0.001$, signed-rank test), SLiR ($p < 0.001$), and SMLR ($p < 0.005$) for all basis shapes.

Additionally, we tested ordinal logistic regression with elastic net regularization (elastic net-OLR; Zou and Hastie, 2005). The elastic net involves regularization using a sum of the L1-norm and the L2-norm of the weight vector as a penalty term, while having two hyperparameters to be manually adjusted. Because the solution of elastic net-OLR cannot be obtained in an analytic form, we minimized the cost function by the iteratively

[2]https://openfmri.org/
http://brainliner.jp/data/brainliner/Visual_Image_Reconstruction

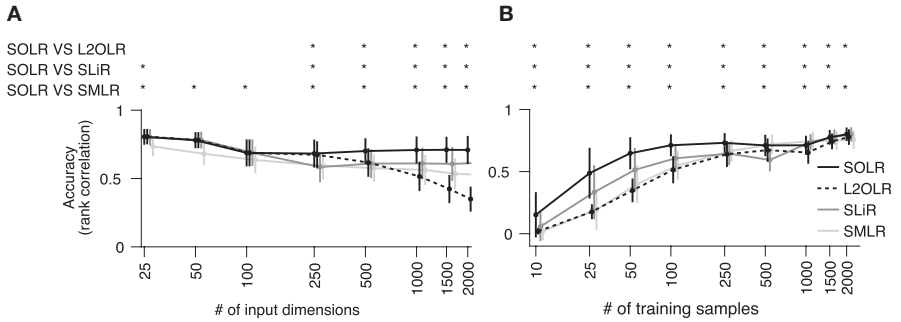

FIGURE 2 | Simulation analysis. **(A)** Decoding accuracy as a function of the number of input dimensions. SOLR, L2OLR, SLiR, and SMLR are trained with 100 samples and tested with 1,000 samples using simulation data. The prediction performance is evaluated by the Spearman rank correlation between true and predicted labels. In this simulation, only 10 input dimensions include information on labels, and the number of irrelevant dimensions is changed to characterize the performance. The mean performance across 100 simulation repetitions is plotted as a function of the total number of input dimensions (*$p < 0.01$, signrank test; error bar: SD across 100 simulation repetitions). **(B)** Decoding accuracy as a function of the number of training data. The mean prediction performance is plotted as a function of the number of training samples (*$p < 0.01$, signrank test; error bar: SD across 100 simulation repetitions). The total number of dimensions is fixed to 1,000.

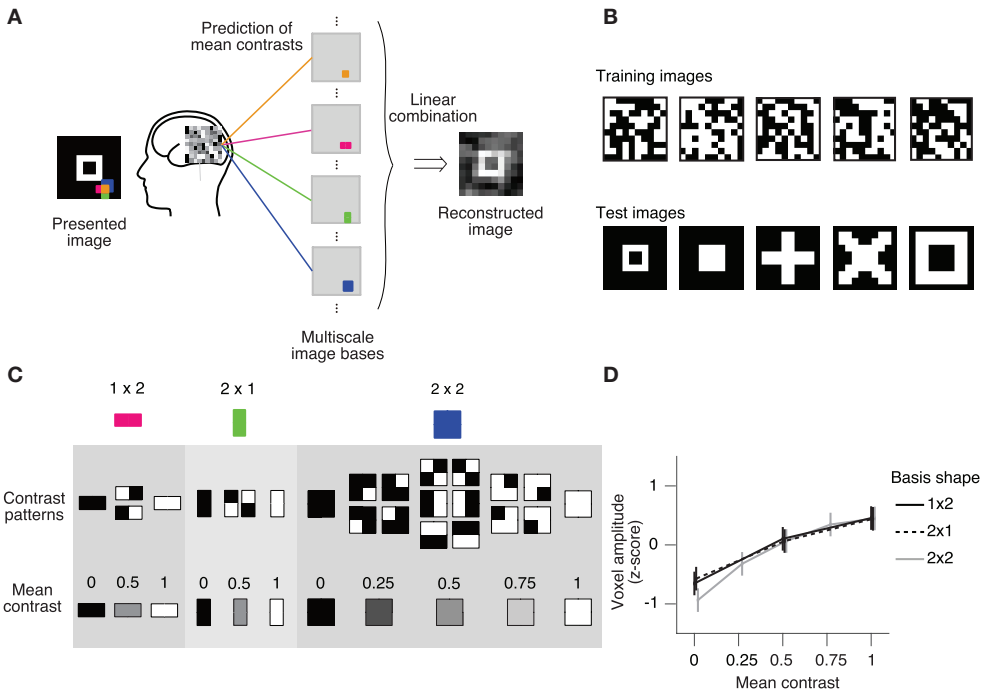

FIGURE 3 | Ordinal regression task with visual image reconstruction data. **(A)** Reconstruction procedure of Miyawaki et al. (2008). A presented image was reconstructed by first predicting the contrasts of image bases from brain activity and then optimally combining the image bases multiplied by the predicted contrasts. Image bases were 1 × 1-, 1 × 2-, 2 × 1-, and 2 × 2-shaped regions covering a 10 × 10-pixel image with overlaps (a total of 361 regions; 1 × 1, 100; 1 × 2, 90; 2 × 1, 90; 2 × 2, 81). **(B)** Examples of stimulus images. Random binary images (upper) and simple geometric shapes (lower) were respectively presented in the random image session and the figure image session of the experiment. In reconstruction analysis, data in the random image session were used to train decoders and data in the figure image session were used as test data. **(C)** Contrast patterns and mean contrast in image bases. 1 × 2- and 2 × 1-shaped regions took one of the four contrast patterns, and the mean contrast within the region took one of the three values (left, middle). 2 × 2-shaped regions took one of the 16 contrast patterns, and the mean contrast took one of the five values. **(D)** Relationship between the voxel amplitude and mean contrast. Voxels responsive to each local region were selected in a univariate analysis of the random image session data. The voxel amplitudes were then averaged across trials for each mean contrast. Error bars show the 95% confidence intervals across local regions.

reweighted shrinkage method (Chartrand and Yin, 2008). This minimization was performed with the same stopping criterion as in SOLR with a step tolerance of 0.001. An elastic net-OLR training with a particular hyperparameter set took about half the time of an SOLR training. Whereas a careful tuning of the hyperparameters with many repetitions could lead to a slightly

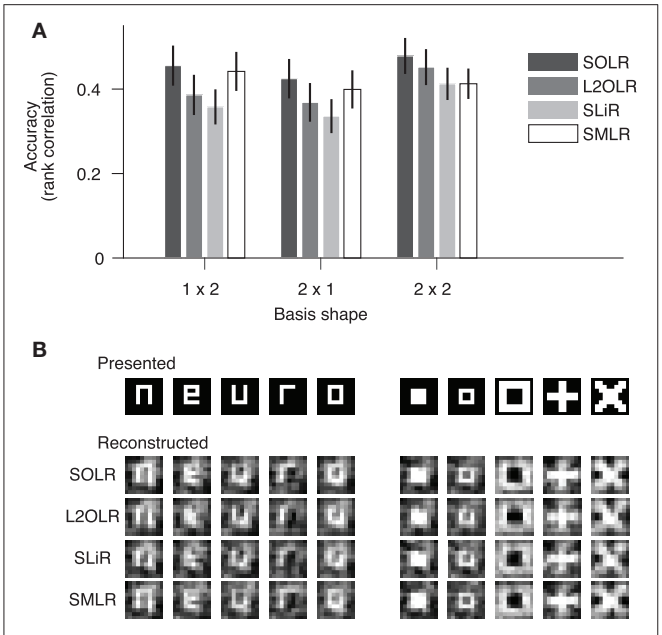

FIGURE 4 | Performance on visual image reconstruction data. **(A)** Prediction of mean contrasts of image bases. For each basis, the mean contrast was predicted by SOLR, L2OLR, SLiR, and SMLR. Performance was evaluated using the Spearman rank correlation between true and predicted mean contrast values across test samples. The mean performance averaged across basis locations is shown. Error bars show the 95% confidence intervals across basis locations. **(B)** Reconstructed images. Presented images and images reconstructed with the four algorithms are shown.

superior or comparable performance to SOLR, the performance of elastic net-OLR fell below that of SOLR in most ranges of hyperparameters. Note that SOLR has no hyperparameters, and does not require a time-consuming tuning.

To confirm that the sparseness introduced in SOLR, SLiR, and SMLR automatically selects a subset of voxels, we counted the number of the voxels with non-zero weights in each local decoder. While SOLR, SLiR, and SMLR selected 88 $\pm$ 15, 180 $\pm$ 7.7, and 56 $\pm$ 17 (mean $\pm$ SD) voxels out of the 996 V1 voxels, respectively, all the weights estimated by L2OLR were nonzero. We also confirmed that the majority of voxels selected by SOLR were located in the hemisphere contralateral to the image basis location, consistent with the known retinotopic mapping.

To examine whether the sparseness in SOLR efficiently prevents overfitting, we conducted an additional analysis where the number of training samples was randomly reduced to 50 samples from all the 352 samples. As the size of the training data becomes smaller, the risk of overfitting increases. In this condition, SOLR outperformed L2OLR with a larger difference ($r = 0.25 \pm 0.17$ for SOLR, 0.15 ± 0.13 for L2OLR; mean $\pm$ SD across randomly selected 50 local image bases) than when all training samples were used ($r = 0.46 \pm 0.19$ for SOLR, 0.41 ± 0.19 for L2OLR).

We finally reconstructed visual images according to the predictions of the models with the same procedure as adopted by Miyawaki et al. (2008), and compared reconstructed images between the four algorithms (**Figure 4B**). The image bases were multiplied by the predicted mean contrasts and then linearly combined with optimized weights to produce a single image (see Miyawaki et al., 2008 for details). Although the differences are not remarkable in visual inspection, the spatial correlations between presented images and images reconstructed with SOLR were higher than those of the other algorithms ($p < 0.05$, signed-rank test).

DISCUSSION

We developed a new algorithm for ordinal variable decoding by combining OLR with Bayesian sparse weight estimation. The proposed algorithm, SOLR, was compared with three other methods: (1) ordinal logistic regression without a sparse constraint, L2OLR; (2) a regression model with the same Bayesian sparse constraint, SLiR; and (3) a classification model with the same Bayesian sparse constraint, SMLR. In analyses using simulation and real fMRI data, SOLR had better prediction performance than the other three methods. These results suggest that SOLR is a useful tool in decoding analyses where the target variable can be regarded as ordinal.

Ordinal variables naturally emerge in decoding analysis; however, they have been predicted using classification models (Miyawaki et al., 2008; Staeren et al., 2009; Baucom et al., 2012; Cortese et al., 2016, 2017) or regression models (Chu et al., 2011; Valente et al., 2011; Nishio et al., 2012; Chang et al., 2015). By the nature of the ordinal variable, the levels an ordinal variable takes have a relative order but the distances between levels are not given. Because regression models use a metric in the label space and their predictions depend on it (**Figure 1A**), regression models are not appropriate for ordinal variable prediction. Meanwhile, classification models do not need the distances between classes. However, the complexity of classification models rapidly grows as the number of classes becomes large, which increases the chance of overfitting (**Figure 1B**). Here, to predict ordinal variables in decoding analysis, we introduced OLR (McCullagh, 1980), one of the known generalized linear models whose output variable is assumed to be an ordinal variable. To prevent overfitting in decoding analysis where a large number of voxels are used as input, we proposed a new method, SOLR, by combining OLR with a Bayesian sparse weight estimation method (MacKay, 1992; Neal, 1996; Yamashita et al., 2008).

In the analysis using simulation data, SOLR outperformed L2OLR, SLiR, and SMLR as the number of input dimensions increased or the number of training data decreased (**Figure 2**). The comparison between SOLR and L2OLR suggests that the sparseness introduced into SOLR prevents overfitting efficiently and improves the decoding performance, which is consistent with the results of previous studies analyzing the utility of the sparseness using classification models (Yamashita et al., 2008; Ryali et al., 2010). A comparison among SOLR, SLiR, and SMLR showed that the appropriate treatment of a given relative order by OLR also leads to better decoding performance.

In analysis using real fMRI data, the same four methods were compared and SOLR had better prediction performance than

L2OLR, SLiR, and SMLR (**Figure 4A**). While the same contrast prediction task was conducted with SMLR in the previous study (Miyawaki et al., 2008), we found that the prediction can be improved by introducing SOLR. Although the resultant reconstructed images of SOLR and SMLR appear similar (**Figure 4B**), SOLR had a slightly higher spatial correlation than SMLR. Note also that SOLR does not require to manually tune hyperparameters as in elastic net regularization. These results suggest that SOLR would work well in a practical situation of fMRI decoding analysis.

Additional analyses showed that SOLR outperformed L2OLR even better when the amount of training data was reduced, and that voxels to which large weights were assigned by SOLR are mainly distributed in the contralateral hemisphere to the image basis locations. These results suggest that SOLR prevents overfitting by selecting physiologically relevant voxels for prediction.

Taken together, SOLR is expected to provide a principled and effective method of decoding ordinal variables. While ordinal variables have been predicted using classification models or regression models in previous decoding studies, we found that SOLR outperformed linear classification and regression models with the same type of sparseness. The results suggest that SOLR would be helpful in decoding analysis where an ordinal variable is used as the target variable and would allow us to better characterize the neural representations of subjective states that are quantified by subjective ratings, such as impressions, emotional feelings, and confidence.

AUTHOR CONTRIBUTIONS

ES, KM, and YK designed the study. KM developed the algorithm. ES and SA performed analyses. ES, KM, and YK wrote the manuscript.

FUNDING

This research was supported by grants from JSPS KAKENHI (grant number JP15H05920, JP15H05710), the ImPACT Program of Council for Science, the Technology and Innovation (Cabinet Office, Government of Japan), the Strategic International Cooperative Program (JST/AMED), AMED under Grant Number JP18dm0107151h, and the New Energy and Industrial Technology Development Organization (NEDO).

ACKNOWLEDGMENTS

The authors thank Mohamed Abdelhack, Tomoyasu Horikawa, Ken Shirakawa, Yu Takagi, and Mitsuaki Tsukamoto, for helpful comments on the manuscript and analysis.

REFERENCES

Attias, H. (1999). "Inferring parameters and structure of latent variable models by variational Bayes," in *Proceedings 15th Conference on Uncertainty in Artificial Intelligence,* (Stockholm: Morgan Kaufmann Pub), 21–30.

Baucom, L. B., Wedell, D. H., Wang, J., Blitzer, D. N., and Shinkareva, S. V. (2012). Decoding the neural representation of affective states. *Neuroimage* 59, 718–727. doi: 10.1016/j.neuroimage.2011.07.037

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. New York, NY: Springer.

Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A., and Wager, T. D. (2015). A sensitive and specific neural signature for picture-induced negative affect. *PLoS Biol.* 13:e1002180. doi: 10.1371/journal.pbio.1002180

Chartrand, R., and Yin, W. (2008). "Iteratively reweighted algorithms for compressive sensing," in *IEEE International Conference on Acoustics, Speech and Signal Processing* (Las Vegas, NV), 3869–3872

Chu, C., Ni, Y., Tan, G., Saunders, C. J., and Ashburner, J. (2011). Kernel regression for fMRI pattern prediction. *Neuroimage* 56, 662–673. doi: 10.1016/j.neuroimage.2010.03.058

Cohen, J. R., Asarnow, R. F., Sabb, F. W., Bilder, R. M., Bookheimer, S. Y., Knowlton, B. J., et al. (2011). Decoding continuous variables from neuroimaging data: basic and clinical applications. *Front. Neurosci.* 5:75. doi: 10.3389/fnins.2011.00075

Cortese, A., Amano, K., Koizumi, A., Kawato, M., and Lau, H. (2016). Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. *Nat. Commun.* 7:1366918. doi: 10.1038/ncomms 13669

Cortese, A., Amano, K., Koizumi, A., Lau, H., and Kawato, M. (2017). Decoded fMRI neurofeedback can induce bidirectional confidence changes within single participants. *NeuroImage* 149, 323–337. doi: 10.1016/j.neuroimage.2017.01.069

Haynes, J. D., and Rees, G. (2006). Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.* 7, 523–534. doi: 10.1038/nrn1931

Kamitani, Y., and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nat. Neurosci.* 8, 679–685. doi: 10.1038/nn1444

Kamitani, Y., and Tong, F. (2006). Decoding seen and attended motion directions from activity in the human visual cortex. *Curr. Biol.* 16, 1096–1102. doi: 10.1016/j.cub.2006.04.003

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Comput.* 4, 415–447. doi: 10.1162/neco.1992.4.3.415

McCullagh, P. (1980). Regression models for ordinal data. *J. R. Stat. Soc. B* 42, 109–142.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M. A., Morito, Y., Tanabe, H. C., et al. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron* 60, 915–929. doi: 10.1016/j.neuron.2008.11.004

Neal, R. M. (1996). "Bayesian Learning for Neural Networks," in *Lecture Notes in Statistics* No. 118. (New York, NY: Springer-Verlag), 16–17.

Nishio, A., Goda, N., and Komatsu, H. (2012). Neural selectivity and representation of gloss in the monkey inferior temporal cortex. *J. Neurosci.* 32, 10780–10793. doi: 10.1523/JNEUROSCI.1095-12.2012

Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* 10, 424–430. doi: 10.1016/j.tics.2006.07.005

Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage* 45, S199–S209. doi: 10.1016/j.neuroimage.2008.11.007

Poldrack, R. A., Barch, D. M., Mitchell, J. P., Wager, T. D., Wagner, A. D., Devlin, J. T., et al. (2013). Toward open sharing of task-based fMRI data: the OpenfMRI project. *Front. Neuroinform.* 7:12. doi: 10.3389/fninf.2013.00012

Ryali, S., Supekar, K., Abrams, D. A., and Menon, V. (2010). Sparse logistic regression for whole-brain classification of fMRI data. *NeuroImage* 51, 752–764. doi: 10.1016/j.neuroimage.2010.02.040

Smith, A., Bernheim, B. D., Camerer, C., and Rangel, A. (2014). Neural activity reveals preferences without choices. *Am. Econ. J. Microecon.* 6, 1–36. doi: 10.1257/mic.6.2.1

Staeren, N., Renvall, H., De Martino, F., Goebel, R., and Formisano, E. (2009). Sound categories are represented as distributed patterns in the human auditory cortex. *Curr. Biol.* 19, 498–502. doi: 10.1016/j.cub.2009.01.066

Takemiya, M., Majima, K., Tsukamoto, M, and Kamitani, Y. (2016). BrainLiner: A Neuroinformatics Platform for Sharing Time-Aligned Brain-Behavior Data. *Front. Neuroinform.* 10:3. doi: 10.3389/fninf.2016.00003

Tipping, M. (2001). Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn Res.* 1, 211–244. doi: 10.1162/15324430152748236

Valente, G., De Martino, F., Esposito, F., Goebel, R., and Formisano, E. (2011). Predicting subject-driven actions and sensory experience in a virtual world with Relevance Vector Machine Regression of fMRI data. *NeuroImage* 56, 651–661. doi: 10.1016/j.neuroimage.2010.09.062

Winship, C., and Mare, R. (1984). Regression models with ordinal variables. *Am. Sociol. Rev*. 49, 512–525. doi: 10.2307/2095465

Yamashita, O., Sato, M. A., Yoshioka, T., Tong, F., and Kamitani, Y. (2008). Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. *NeuroImage* 42, 1414–1429. doi: 10.1016/j.neuroimage.2008.05.050

Zou, H., and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc.* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

We estimated the MAP solution of the weight and threshold parameters in SOLR using the mean-field variational Bayesian approximation and Laplace approximation. In this estimation procedure, we approximate the posterior distribution as

$$p\left(\mathbf{w},\boldsymbol{\mu},\boldsymbol{\alpha}|\mathbf{X},\mathbf{Y}\right)\approx q_1\left(\mathbf{w},\boldsymbol{\mu}\right)q_2\left(\boldsymbol{\alpha}\right). \tag{12}$$

$q_1$and q_2 are probability density functions that are iteratively updated to obtain a better approximation. In the variational Bayesian method, $q_1$and q_2 are alternately updated using the rules

$$\log q_1\left(\mathbf{w},\boldsymbol{\mu}\right):=\langle\log p(\mathbf{w},\boldsymbol{\mu},\boldsymbol{\alpha},\mathbf{Y}|\mathbf{X})\rangle_{q_2(\boldsymbol{\alpha})}+\text{const} \tag{13}$$

and

$$\log q_2\left(\boldsymbol{\alpha}\right):=\langle\log p(\mathbf{w},\boldsymbol{\mu},\boldsymbol{\alpha},\mathbf{Y}|\mathbf{X})\rangle_{q_1(\mathbf{w},\boldsymbol{\mu})}+\text{const}, \tag{14}$$

where $\langle x\rangle_{q(x)}$ denotes the expectation of x with respect to the probability distribution $q\left(x\right)$. Each update decreases the Kullback–Leibler divergence between $p\left(\mathbf{w},\boldsymbol{\mu},\boldsymbol{\alpha}|\mathbf{X},\mathbf{Y}\right)$ and $q_1\left(\mathbf{w},\boldsymbol{\mu}\right)q_2\left(\boldsymbol{\alpha}\right)$, which makes $q_1\left(\mathbf{w},\boldsymbol{\mu}\right)q_2\left(\boldsymbol{\alpha}\right)$ a better approximation of the posterior distribution (Attias, 1999; Bishop, 2006; Yamashita et al., 2008). In the following, we describe the procedure of updating q_1 and q_2, respectively.

To update q_1, the right side of (13) is rewritten as

$$\sum_{n=1}^{N}\sum_{c=1}^{C}y_{nc}\log\left[F\left(\mu_c-\mathbf{w}^{\mathrm{T}}\mathbf{x}_n\right)-F\left(\mu_{c-1}-\mathbf{w}^{\mathrm{T}}\mathbf{x}_n\right)\right]$$
$$-\frac{1}{2}\sum_{d=1}^{D}\bar{\alpha}_d{w_d}^2+const, \tag{15}$$

where $\bar{\alpha}_d=\langle\alpha_d\rangle_{q_2(\alpha)}$. F is the logistic sigmoid function whose definition was given in (4). The probability distribution function whose logarithm is given by (15) cannot be obtained in analytic form, and we therefore applied the Laplace approximation. In the approximation, (15) is replaced with its second-order Taylor series expansion around the maximum, and the update equation of q_1 is rewritten as

$$q_1\left(\mathbf{w},\boldsymbol{\mu}\right):=\varphi\left(\mathbf{w};\bar{\mathbf{w}},\Sigma_{\mathbf{w}}\right)\varphi\left(\boldsymbol{\mu};\bar{\boldsymbol{\mu}},\Sigma_{\boldsymbol{\mu}}\right), \tag{16}$$

where, the function $\varphi(\cdot;\mathbf{m},\Sigma)$ denotes the probability density function of the multidimensional Gaussian distribution with mean $\mathbf{m}$ and covariance Σ. $\bar{\mathbf{w}}$ and $\bar{\boldsymbol{\mu}}$ are the values of $\mathbf{w}$ and $\boldsymbol{\mu}$ that maximize $\langle\log p(\mathbf{w},\boldsymbol{\mu},\boldsymbol{\alpha},\mathbf{Y}|\mathbf{X})_{q_2(\boldsymbol{\alpha})}\rangle$. $\Sigma_{\mathbf{w}}$ is the Hessian matrix of $\langle\log p(\mathbf{w},\ \boldsymbol{\mu},\boldsymbol{\alpha},\mathbf{Y}|\mathbf{X})_{q_2(\boldsymbol{\alpha})}\rangle$ with respect to $\mathbf{w}$ at $\bar{\mathbf{w}}$. $\Sigma_{\boldsymbol{\mu}}$ is the Hessian matrix of $\langle\log p(\mathbf{w},\ \boldsymbol{\mu},\boldsymbol{\alpha},\mathbf{Y}|\mathbf{X})_{q_2(\boldsymbol{\alpha})}\rangle$ with respect to $\boldsymbol{\mu}$ at $\bar{\boldsymbol{\mu}}$. The calculation of $\bar{\mathbf{w}}$ and $\bar{\boldsymbol{\mu}}$ was performed using a gradient method in the present study. As $\Sigma_{\boldsymbol{\mu}}$ is not used in the update of q_2, only $\bar{\mathbf{w}}$, $\bar{\boldsymbol{\mu}}$and $\Sigma_{\mathbf{w}}$ were calculated in the present study.

In the update of q_2, using the relationship (16), the update rule is given by

$$q_2\left(\boldsymbol{\alpha}\right):=\prod_{d=1}^{D}\psi\left(\alpha_d;\bar{\alpha}_d,\gamma_d\right), \tag{17}$$

where, $\varphi(\cdot;\mathbf{m},\Sigma)$ denotes the probability density function of the gamma distribution with mean $\bar{\alpha}$ and degree of freedom γ . γ_d is 0.5 regardless of d, and $\bar{\alpha}_d$ is given by

$$\bar{\alpha}_d:=\frac{1}{\bar{w}_d^2+\Sigma_{\mathbf{w}(d,d)}}, \tag{18}$$

where $\bar{w}_d$ is the d-th element of $\bar{\mathbf{w}}$ and $\Sigma_{\mathbf{w}(d,d)}$ is the d-th diagonal element of $\Sigma_{\mathbf{w}}$. To accelerate convergence, a modified rule adopted in previous studies (MacKay, 1992; Yamashita et al., 2008) was used instead of (18):

$$\bar{\alpha}_d:=\frac{1-\bar{\alpha}_d\Sigma_{\mathbf{w}(d,d)}}{\bar{w}_d^2}. \tag{19}$$

As initial parameters, $\bar{\alpha}_1,\cdots,\bar{\alpha}_D$ were set to 1, and q_1 and q_2 were alternately updated 100 times in this study.