

Sparse Bayesian Correntropy Learning for Robust Muscle Activity Reconstruction from Noisy Brain Recordings

Yuanhao Li, Badong Chen, *Senior Member, IEEE*,
Natsue Yoshimura, Yasuharu Koike, and Okito Yamashita

Abstract—Sparse Bayesian learning has promoted many effective frameworks for brain activity decoding, especially for the reconstruction of muscle activity. However, existing sparse Bayesian learning mainly employs Gaussian distribution as error assumption in the reconstruction task, which is not necessarily the truth in the real-world application. On the other hand, brain recording is known to be highly noisy and contains many non-Gaussian noises, which could lead to significant performance degradation for sparse Bayesian learning method. The goal of this paper is to propose a new robust implementation for sparse Bayesian learning, so that robustness and sparseness can be realized simultaneously. Motivated by the great robustness of maximum correntropy criterion (MCC), we proposed an integration of MCC into the sparse Bayesian learning regime. To be specific, we derived the explicit error assumption inherent in the MCC and then leveraged it for the likelihood function. Meanwhile, we used the automatic relevance determination (ARD) technique for the sparse prior distribution. To fully evaluate the proposed method, a synthetic dataset and a real-world muscle activity reconstruction task with two different brain modalities were employed. Experimental results showed that our proposed sparse Bayesian correntropy learning framework improves significantly the robustness in a noisy regression task. The proposed method can realize higher correlation coefficient and lower root mean squared error in the real-world muscle activity reconstruction tasks. Sparse Bayesian correntropy learning provides a powerful tool for neural decoding which can promote the development of brain-computer interfaces.

Index Terms—brain decoding, sparse Bayesian learning, maximum correntropy criterion, muscle activity reconstruction, robustness.

I. INTRODUCTION

BRAIN-computer interface (BCI) has been revealing hopeful and promising prospects for paralyzed and amputated people towards their rehabilitation and communication [1], [2]. In particular, previous studies have demonstrated the capability of BCI to reproduce the kinetic functions [3], [4], which would help neurological recovery for post-stroke motor rehabilitation and develop the assistive prosthesis for motor-impaired people. The crucial procedure for a motion-related BCI is to accurately decode the movements from simultaneous brain activity, which is usually realized by two different strategies. The first scheme is motor imagery [5], a classification scenario with pre-defined labels. However, the discrete label of motor imagery may result in an inadequate capability for the reproduction of complicated motion with only the predetermined motion intention. Another strategy is to reconstruct the continuous movements, which can reflect the motion in real time and thus exhibits larger potential to advance the development of motion-related BCI system [4].

Electromyography (EMG) signal directly reflects the muscle activation, which comprises the information for related motion [6], and can be employed to restore the torques and trajectories of muscle activity [7], and to control a prosthetic hand [8], [9]. Previous studies have investigated the decoding of EMG-based muscle activity by different brain recording methods [10]–[14], including the widely utilized electroencephalogram (EEG) and functional magnetic resonance imaging (fMRI). Further, some have employed the multimodal source imaging methods which could merge the merits of EEG and fMRI, thus realizing better motion decoding performance by the estimated cortical current sources [15]–[19].

The regression algorithm plays a vital role for reconstructing muscle activities, which predicts the continuous response from brain recordings. Although recent advances have utilized deep learning methods for brain activity decoding [20], conventional linear regression models are still popularly utilized for motion decoding from neural signals in recent studies, e.g., [11], [13], [19], which is also mentioned in a recent review paper [4]. This is mainly because the deep learning methods need copious data for training while collecting a large amount of brain recordings

This work was supported in part by Japan Society for the Promotion of Science (JSPS) KAKENHI under Grants 19H05728, 20H00600, and 23H03433, in part by Innovative Science and Technology Initiative for Security under Grant JPJ004596 ATLA, in part by Moonshot Program 9 under Grant JPMJMS2291, and in part by the National Natural Science Foundation of China under Grants U21A20485 and 62311540022. (Corresponding author: Yuanhao Li.)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of National Center of Neurology and Psychiatry.

Yuanhao Li and Okito Yamashita are with the Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan, and also with the Brain Information Communication Research Laboratory Group, Advanced Telecommunications Research Institute International, Kyoto 619-0237, Japan. (correspondence e-mail: yuanhao.li@riken.jp)

Badong Chen is with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China.

Natsue Yoshimura is with the School of Computing, Tokyo Institute of Technology, Yokohama 226-8503, Japan.

Yasuharu Koike is with the Institute of Innovative Research, Tokyo Institute of Technology, Yokohama 226-8503, Japan.

The code is available at <https://sites.google.com/view/liyuanhao/code>.

is not always feasible. Moreover, linear model can improve the interpretability for the neurological patterns [12]. In particular, to alleviate the overfitting problem with a small dataset, sparse Bayesian learning has been widely leveraged for brain activity analysis [11], [14]–[18], [21]–[23], which utilizes the sparsity-promoting prior distribution on model parameter and calculates the maximum a posteriori (MAP) estimation. Thus, the model training and covariate selection can be realized simultaneously.

Another significant problem for neural decoding tasks is that brain recordings are extremely noisy and contain non-Gaussian noises [24], [25], conflicting with the traditional Gaussian error assumption utilized in sparse Bayesian learning. Consequently, this may cause a significant performance degradation for brain activity decoding. Although preprocessing methods have been developed to denoise the brain recordings, such as independent component analysis (ICA) [26], it is still hard to guarantee that all recorded noise can be removed favorably by preprocessing.

The goal of this paper is to propose a robust implementation of sparse Bayesian learning, so that the covariate selection can be realized with alleviating the negative effect of the recording noise. The motivation of this research is maximum correntropy criterion (MCC), which was proposed in information theoretic learning (ITL) [27] and has achieved extraordinary robustness in many machine learning scenarios, including regression [28], [29], classification [30]–[32], dimensionality reduction [33], etc. This study aims to integrate MCC into the sparse Bayesian learning framework so that sparseness and robustness could be realized simultaneously. This proposal was proved effective for classification tasks by our previous study [32], and the present paper aims to extend the proposed sparse Bayesian correntropy learning framework into the reconstruction (regression) scene. Main contributions of the present paper are outlined as follows:

1. Our preliminary work [32] lacked an interpretation of this proposal, i.e., using MCC in the sparse Bayesian learning. This paper aims to provide an insightful interpretation for this integration, by deriving the error assumption in MCC.
2. MCC-based robust likelihood function is utilized with the automatic relevance determination (ARD) prior technique for robust sparse regression with only one hyperparameter to be tuned.
3. While the conference paper for the present study [34] has made the theoretical advancements, it lacked a systematic validation of the proposed method in the real-world neural decoding applications, which only used a toy dataset with a noisy output. To fully demonstrate the practical impacts of our method for neural decoding, the present paper aims at a comprehensive evaluation, which includes a synthetic dataset with a noisy input and a real-world muscle activity reconstruction task using cortical current source and EEG, respectively.

II. METHOD

A. Problem Setting

Regression aims at learning a model from the dataset which can realize a prediction for continuous response variables. One can consider the canonical linear regression model as follows:

$$t = \mathbf{x}\mathbf{w} + \epsilon \quad (1)$$

where t is the model output, $\mathbf{x} = (x_1, x_2, \dots, x_D) \in \mathbb{R}^{1 \times D}$ is a D -dimensional covariate, $\mathbf{w} = (w_1, w_2, \dots, w_D)^T \in \mathbb{R}^{D \times 1}$ is the model parameter, while ϵ denotes the error term. Usually, one has a finite dataset $\{(\mathbf{x}_n, t_n)\}_{n=1}^N$, with N samples to train the regression model. The dataset can be denoted by $(\mathbf{X}, \mathbf{t})$, in which $\mathbf{t} = (t_1, t_2, \dots, t_N)^T \in \mathbb{R}^{N \times 1}$ contains all the response variables, and $\mathbf{X} \in \mathbb{R}^{N \times D}$ denotes the collection of covariates, where each row is an individual covariate $\mathbf{x}_n$.

B. Conventional Sparse Bayesian Learning with ARD

To construct the relationship between covariate $\mathbf{x}$ and output t , Bayesian learning methods first need to make an assumption for the error term ϵ . The most common assumption for ϵ is to leverage the zero-mean Gaussian distribution with the variance σ^2 . Thus, the probability density function (PDF) of t could be written by $p(t|\mathbf{x}) = \mathcal{N}(t|\mathbf{x}\mathbf{w}, \sigma^2)$, which denotes the Gaussian distribution over t with mean value $\mathbf{x}\mathbf{w}$ and variance σ^2 . Then the likelihood is written by assuming the sample independence

$$p(\mathbf{t}|\mathbf{w}, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{t} - \mathbf{X}\mathbf{w}\|^2\right\} \quad (2)$$

Maximum likelihood estimation (MLE) could give the solution that maximizes the probability of occurrence for the data under the corresponding error assumption. The Gaussian assumption based MLE is also equivalent to the least-square (LS) criterion and the mean squared error (MSE) loss function.

To acquire a sparse solution, one can further use a sparsity-promoting prior distribution on the model parameter, and then calculate the MAP estimation. For example, one could assume $\mathbf{w}$ obeys a Laplace distribution, which is equal to L_1 -regularization. Here, we employ the automatic relevance determination (ARD), since it can exclude the hyperparameter which controls the sparseness. ARD is a hierarchical prior technique that first assumes zero-mean Gaussian distribution for each entry of the model parameter with individual variances

$$p(\mathbf{w}|\mathbf{a}) = \prod_{d=1}^D p(w_d|a_d) = \prod_{d=1}^D \mathcal{N}(w_d|0, a_d^{-1}) \quad (3)$$

where a_d is the inverse variance (called relevance parameter) for the distribution of w_d , and $\mathbf{a} = (a_1, a_2, \dots, a_D)$. One can find that a_d regulates the distribution range for corresponding w_d . Each a_d is further assigned with the non-informative prior

$$p(\mathbf{a}) = \prod_{d=1}^D p(a_d) = \prod_{d=1}^D a_d^{-1} \quad (4)$$

This is essentially an improper prior since its integral is infinite and thus can not be normalized [35]. The error variance is also assumed with the non-informative prior

$$p(\sigma^2) = (\sigma^2)^{-1} \quad (5)$$

By defining the likelihood function and the prior distributions, the posterior distribution of all unknown variables is calculated

$$p(\mathbf{w}, \mathbf{a}, \sigma^2|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{w}, \mathbf{a}, \sigma^2)p(\mathbf{w}, \mathbf{a}, \sigma^2)}{p(\mathbf{t})} \quad (6)$$

However, because the normalizing integral for the denominator $p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w}, \mathbf{a}, \sigma^2)p(\mathbf{w}, \mathbf{a}, \sigma^2)d\mathbf{w}d\mathbf{a}d\sigma^2$ is intractable, one

needs approximation for the posterior distribution to obtain the MAP estimation. This issue can be effectively solved by type-II maximum likelihood [36], [37], expectation maximum (EM) [38], and variational inference [39]. During the model training, some a_d will become tremendously large, which indicates that the corresponding w_d is tightly distributed around zero. Hence, these dimensions could be removed from the data. In practice, one usually uses a threshold, and prune the features when their corresponding a_d exceeds the threshold. Because this threshold does not make an obvious influence on the sparseness provided it is set large enough (e.g., 10^6 or 10^8), ARD does not require the hyper-parameter tuning, thus realizing adaptive sparseness.

C. Maximum Correntropy Criterion

The key concept of MCC is correntropy which was proposed as a generalized correlation function for stochastic process [40] and was further extended as a similarity measure between two random variables [28]. Specifically, given the random variables a and b accompanied with their N observations $\{(a_n, b_n)\}_{n=1}^N$, the correntropy between a and b is defined by

$$V(a, b) \triangleq \mathbb{E}_{p(a, b)} [k(a, b)] \quad (7)$$

which is the expectation of the kernel function $k(\cdot, \cdot)$ between a and b with respect to their joint distribution $p(a, b)$. Usually, the kernel function is practiced by the Gaussian kernel function and the empirical estimation of $V(a, b)$ is obtained by

$$\begin{aligned} \hat{V}(a, b) &= \frac{1}{N} \sum_{n=1}^N k_h(a_n, b_n) \\ &= \frac{1}{N} \sum_{n=1}^N \exp\left(-\frac{(a_n - b_n)^2}{2h}\right) \end{aligned} \quad (8)$$

in which $k_h(a, b) \triangleq \exp(-(a - b)^2/2h)$ denotes the Gaussian kernel function with the kernel bandwidth $h > 0$.

Correntropy provides an approach for robust data characterization, since the correntropy value is mainly determined along $a = b$ by the Gaussian kernel function $k_h(a, b)$, which exhibits a local property and thus suppresses the negative effect of large deviation that is caused by adverse noise. Correntropy extracts more information from the dataset than conventional 2nd-order statistics since it contains all the even-numbered moments (by Taylor expansion on the kernel function). Thus, one can utilize correntropy as the objective function in the supervised machine learning to maximize the similarity between the prediction and desired target, called maximum correntropy criterion (MCC).

D. Error Assumption in MCC

To integrate MCC into the sparse Bayesian learning context with theoretical interpretation, we first concentrate on deriving the explicit error assumption inherent in the operation of MCC. MCC-based linear regression, named as maximum correntropy regression (MCR), can be expressed by the following objective

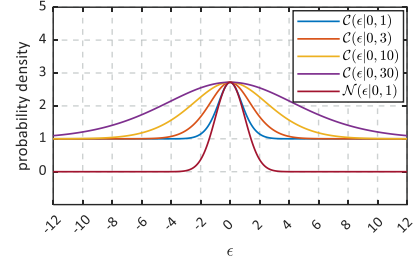


Fig. 1: $\mathcal{C}(\epsilon|0, h)$ under different h values compared to standard Gaussian distribution $\mathcal{N}(\epsilon|0, 1)$. Gaussian distribution is scaled to show the same maximum value as $\mathcal{C}(\epsilon|0, h)$ for comparison.

function:

$$\begin{aligned} \mathbf{w} &= \arg \max_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^N \exp\left(-\frac{(t_n - \mathbf{x}_n \mathbf{w})^2}{2h}\right) \\ &= \arg \max_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^N \exp\left(-\frac{\epsilon_n^2}{2h}\right) \end{aligned} \quad (9)$$

where $\epsilon_n \triangleq t_n - \mathbf{x}_n \mathbf{w}$ denotes the n -th prediction error. Given a fixed dataset $\{(\mathbf{x}_n, t_n)\}_{n=1}^N$, the denominator N is a constant in the objective function and thus could be omitted. As a result, (9) could be transformed to a multiplication form by utilizing an extra exponential function

$$\begin{aligned} \mathbf{w} &= \arg \max_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^N \exp\left(-\frac{\epsilon_n^2}{2h}\right) \\ &= \arg \max_{\mathbf{w}} \sum_{n=1}^N \exp\left(-\frac{\epsilon_n^2}{2h}\right) \\ &= \arg \max_{\mathbf{w}} \prod_{n=1}^N \exp\left\{\exp\left(-\frac{\epsilon_n^2}{2h}\right)\right\} \end{aligned} \quad (10)$$

This step is simple but enlightening that transfers the empirical estimation of correntropy to be an accumulative multiplication. It provides a new perspective to interpret MCC as a likelihood function. One can find extraordinarily that (10) can be regarded as a novel MLE if the following error assumption $\mathcal{C}$ is utilized

$$\mathcal{C}(\epsilon|0, h) \triangleq \exp\left\{\exp\left(-\frac{\epsilon^2}{2h}\right)\right\} \quad (11)$$

Here, this novel error assumption $\mathcal{C}$ is named as a correntropy-aware PDF. $\mathcal{C}(\epsilon|0, h)$ denotes the distribution for ϵ with a zero mean and a shape parameter h . Apparently, the integral of this novel PDF is infinite. Further, it is even more improper because when ϵ becomes arbitrarily large, the probability density would be close to 1 instead of the normal case 0. Hence, this proposed PDF $\mathcal{C}$ is called as a deviant error assumption. Fig. 1 provides some examples for $\mathcal{C}(\epsilon|0, h)$ under different h values compared to scaled Gaussian distribution (with a same maximum value).

E. Maximum Correntropy Regression with ARD

By exposing the inherent error assumption in MCR, one can then formally formulate a robust likelihood function motivated by the MCC. Using the error assumption $\mathcal{C}(\epsilon|0, h)$, in the linear regression model, the distribution of t is $p(t|\mathbf{x}) = \mathcal{C}(t|\mathbf{x}\mathbf{w}, h)$.

By assuming the independence between each sample, one will obtain the following robust likelihood function

$$\begin{aligned} p(\mathbf{t}|\mathbf{w}, h) &= \prod_{n=1}^N \mathcal{C}(t_n|\mathbf{x}_n\mathbf{w}, h) \\ &= \prod_{n=1}^N \exp\left\{\exp\left(-\frac{(t_n - \mathbf{x}_n\mathbf{w})^2}{2h}\right)\right\} \end{aligned} \quad (12)$$

One can observe that the maximum for this likelihood function is exactly equal to the original objective function of MCC (9).

In what follows, we focus on deriving the integration of the MCC-based likelihood function with ARD technique to realize robust sparse regression from the Bayesian learning viewpoint. Using the robust likelihood (12), however, results in a problem that the posterior distribution over $\mathbf{w}$ cannot be obtained analytically, since likelihood and prior are not conjugated as before. To address this issue, the variational inference [35] is leveraged to approximate the posterior distributions. For simplicity, first, one can consider the shape parameter h (i.e. kernel bandwidth) as a fixed parameter.

Variational inference approximates the true posterior distribution through a surrogate function. To be specific, to estimate and maximize the posterior distribution $p(\mathbf{w}, \mathbf{a}|\mathbf{t}, h)$, the proxy distribution $q(\mathbf{w}, \mathbf{a})$ is introduced. To maximize the similarity between $p(\mathbf{w}, \mathbf{a}|\mathbf{t}, h)$ and $q(\mathbf{w}, \mathbf{a})$, the usual way is minimizing their Kullback-Leibler divergence. Equivalently, one could also minimize the following free energy function [35]

$$\begin{aligned} F(q(\mathbf{w}, \mathbf{a})) &\triangleq -\mathbb{E}_{q(\mathbf{w}, \mathbf{a})} \left[\log \frac{p(\mathbf{w}, \mathbf{a}, \mathbf{t}, h)}{q(\mathbf{w}, \mathbf{a})} \right] \\ &= -\int q(\mathbf{w}, \mathbf{a}) \log \frac{p(\mathbf{w}, \mathbf{a}, \mathbf{t}, h)}{q(\mathbf{w}, \mathbf{a})} d\mathbf{w} d\mathbf{a} \end{aligned} \quad (13)$$

where $p(\mathbf{w}, \mathbf{a}, \mathbf{t}, h)$ denotes the joint distribution. Furthermore, one could apply an extra factorization on the proxy distribution $q(\mathbf{w}, \mathbf{a}) = q_{\mathbf{w}}(\mathbf{w})q_{\mathbf{a}}(\mathbf{a})$. Thus, one has the following free energy

$$\begin{aligned} F(q_{\mathbf{w}}(\mathbf{w})q_{\mathbf{a}}(\mathbf{a})) &= \\ &= -\int q_{\mathbf{w}}(\mathbf{w})q_{\mathbf{a}}(\mathbf{a}) \log \frac{p(\mathbf{w}, \mathbf{a}, \mathbf{t}, h)}{q_{\mathbf{w}}(\mathbf{w})q_{\mathbf{a}}(\mathbf{a})} d\mathbf{w} d\mathbf{a} \end{aligned} \quad (14)$$

Consequently, the free energy could be minimized with respect to $q_{\mathbf{w}}(\mathbf{w})$ and $q_{\mathbf{a}}(\mathbf{a})$ alternately by

$$\begin{aligned} w\text{-step: } \log q_{\mathbf{w}}(\mathbf{w}) &= \mathbb{E}_{q_{\mathbf{a}}(\mathbf{a})} [\log p(\mathbf{w}, \mathbf{a}, \mathbf{t}, h)] + \text{const} \\ a\text{-step: } \log q_{\mathbf{a}}(\mathbf{a}) &= \mathbb{E}_{q_{\mathbf{w}}(\mathbf{w})} [\log p(\mathbf{w}, \mathbf{a}, \mathbf{t}, h)] + \text{const} \end{aligned} \quad (15)$$

in which the logarithm of the joint distribution is

$$\begin{aligned} \log p(\mathbf{w}, \mathbf{a}, \mathbf{t}, h) &= \log p(\mathbf{t}|\mathbf{w}, h) + \log p(\mathbf{w}|\mathbf{a}) + \log p(\mathbf{a}) \\ &= \sum_{n=1}^N \exp\left(-\frac{(t_n - \mathbf{x}_n\mathbf{w})^2}{2h}\right) - \frac{1}{2}\mathbf{w}^T \mathbf{A} \mathbf{w} - \frac{1}{2} \log |\mathbf{A}| + \text{const} \end{aligned} \quad (16)$$

where $\mathbf{A}$ is a $D \times D$ diagonal matrix with the diagonal element

$(a_1, a_2, \dots, a_D)$. Then one has

$$\begin{aligned} \log q_{\mathbf{w}}(\mathbf{w}) &= \sum_{n=1}^N \exp\left(-\frac{(t_n - \mathbf{x}_n\mathbf{w})^2}{2h}\right) - \frac{1}{2}\mathbf{w}^T \mathbb{E}_{q_{\mathbf{a}}(\mathbf{a})} [\mathbf{A}] \mathbf{w} \\ \log q_{\mathbf{a}}(\mathbf{a}) &= -\frac{1}{2} \sum_{d=1}^D a_d \mathbb{E}_{q_{\mathbf{w}}(\mathbf{w})} [w_d^2] - \frac{1}{2} \sum_{d=1}^D \log a_d \end{aligned} \quad (17)$$

To obtain the expectation $\mathbb{E}_{q_{\mathbf{w}}(\mathbf{w})} [w_d^2]$, one needs to formulate the distribution for $q_{\mathbf{w}}(\mathbf{w})$, which still cannot be clarified with an analytical form. Therefore, the Laplacian approximation is applied to $\log q_{\mathbf{w}}(\mathbf{w})$ by a quadratic form

$$\log q_{\mathbf{w}}(\mathbf{w}) \approx \log q_{\mathbf{w}}(\mathbf{w}^*) - \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^T \mathbf{H}(\mathbf{w}^*)(\mathbf{w} - \mathbf{w}^*) \quad (18)$$

where $\mathbf{w}^*$ indicates the value that maximizes $\log q_{\mathbf{w}}(\mathbf{w})$, while $\mathbf{H}(\mathbf{w}^*)$ is the negative Hessian matrix of $\log q_{\mathbf{w}}(\mathbf{w})$ at $\mathbf{w}^*$

$$\begin{aligned} \mathbf{H}(\mathbf{w}) &= -\frac{\partial^2 \log q_{\mathbf{w}}(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^T} \\ &= -\frac{1}{h} \sum_{n=1}^N \mathbf{x}_n^T \left\{ \exp\left(-\frac{\epsilon_n^2}{2h}\right) \left(\frac{\epsilon_n^2}{h} - 1\right) \right\} \mathbf{x}_n + \mathbb{E}_{q_{\mathbf{a}}(\mathbf{a})} [\mathbf{A}] \end{aligned} \quad (19)$$

where ϵ_n is the prediction error produced from the current $\mathbf{w}$. As a result, $q_{\mathbf{w}}(\mathbf{w})$ is approximated by a Gaussian distribution $q_{\mathbf{w}}(\mathbf{w}) \approx \mathcal{N}(\mathbf{w}|\mathbf{w}^*, \mathbf{H}(\mathbf{w}^*)^{-1})$. The expectation $\mathbb{E}_{q_{\mathbf{w}}(\mathbf{w})} [w_d^2]$ is calculated by

$$\mathbb{E}_{q_{\mathbf{w}}(\mathbf{w})} [w_d^2] = w_d^{*2} + s_d^2 \quad (20)$$

where s_d^2 is the d -th diagonal element in $\mathbf{H}(\mathbf{w}^*)^{-1}$. To obtain the maximum solution $\mathbf{w}^*$ for $\log q_{\mathbf{w}}(\mathbf{w})$, one can observe that this is in essence an L_2 -regularized MCR and $\mathbb{E}_{q_{\mathbf{a}}(\mathbf{a})} [\mathbf{A}]$ plays the role of L_2 -regularization weight. The following fixed-point method provides an effective way to optimize $\mathbf{w}$ with very fast convergence [41]

$$\mathbf{w} = (\mathbf{X}^T \mathbf{\Psi} \mathbf{X} + \mathbb{E}_{q_{\mathbf{a}}(\mathbf{a})} [\mathbf{A}])^{-1} \mathbf{X}^T \mathbf{\Psi} \mathbf{t} \quad (21)$$

in which $\mathbf{\Psi}$ denotes a $N \times N$ diagonal matrix, and its diagonal element is $\Psi_{nn} = \exp(-\epsilon_n^2/2h)$.

On the other hand, $q_{\mathbf{a}}(\mathbf{a})$ is analytically equal to the Gamma distribution as follows by applying an exponential function on $\log q_{\mathbf{a}}(\mathbf{a})$ in (17)

$$q_{\mathbf{a}}(\mathbf{a}) = \prod_{d=1}^D q_{a_d}(a_d) = \prod_{d=1}^D \Gamma(a_d|a_d^*, \frac{1}{2}) \quad (22)$$

where $\Gamma(a_d|a_d^*, \frac{1}{2})$ is the Gamma distribution over a_d with the expectation a_d^* and the degree of freedom $\frac{1}{2}$. The expectation a_d^* is computed by

$$a_d^* = \frac{1}{w_d^{*2} + s_d^2} \quad (23)$$

which can also be updated by the following equation for faster convergence [37], [42]

$$a_d^* = \frac{1 - a_d^* s_d^2}{w_d^{*2}} \quad (24)$$

The updated a_d is substituted into $\mathbb{E}_{q_{\mathbf{a}}(\mathbf{a})}[\mathbf{A}]$ for w -step in the next iteration. By optimizing w -step and a -step alternately, the free energy will be minimized, while the MAP estimations for $\mathbf{w}$ and $\mathbf{a}$ will be also obtained, respectively. Algorithm 1 gives a summary for the proposed method that integrates MCR with ARD in a Bayesian framework.

Algorithm 1 *MCR-ARD for Robust Sparse Regression*

- 1: **input:**
 training samples $\{(\mathbf{x}_n, t_n)\}_{n=1}^N$;
 kernel bandwidth h ;
 threshold for covariate pruning $a_{\max}$;
 - 2: **initialize:**
 model parameter w_d ($d = 1, \dots, D$);
 relevance parameter $a_d = 1$ ($d = 1, \dots, D$);
 - 3: **output:**
 model parameter w_d ($d = 1, \dots, D$)
 - 4: **repeat**
 - 5: w -step: update $\mathbf{w}$ according to (21);
 - 6: a -step: update $\mathbf{a}$ according to (24);
 - 7: **if** $a_d \geq a_{\max}$ **then**
 - 8: fix the corresponding w_d to be zero and prune this dimension in the following optimization
 - 9: **end if**
 - 10: **until** the parameter change is small enough or the iteration number is larger than a predetermined limit
-

F. Performance Evaluation

Three performance evaluation scenarios were utilized in this study to fully validate our proposed MCR-ARD algorithm. The three evaluation datasets include a synthetic toy with corrupted covariate, a real-world muscle activity reconstruction task from estimated cortical current source, and the same muscle activity reconstruction task utilizing the EEG recording. The latter two were proposed in our previous study [16], which employed the conventional LSR-ARD to achieve the brain activity decoding.

For performance evaluation, MCR-ARD was compared with its baseline, the conventional sparse Bayesian regression which employs the Gaussian likelihood (least-square criterion), called LSR-ARD, as introduced in Section II-B. Both two algorithms used an identical pruning threshold $a_{\max}$ of 10^6 . The maximal iteration number was set to be 500 for both algorithms. Section IV provides a discussion about the relationship between MCR-ARD and other existing methods.

With regard to the performance indicators for evaluation, the two most commonly used ones in regression were utilized, i.e., correlation coefficient = $Cov(\hat{\mathbf{t}}, \mathbf{t}) / \sqrt{Var(\hat{\mathbf{t}})Var(\mathbf{t})}$, and root mean squared error = $\sqrt{\|\hat{\mathbf{t}} - \mathbf{t}\|^2 / N}$, for which $\hat{\mathbf{t}}$ indicates the model prediction for $\mathbf{t}$, while $Cov(\cdot, \cdot)$ and $Var(\cdot)$ denote the covariance and variance, respectively.

1) Synthetic Dataset:

A demo concerning this synthetic dataset is provided¹ which could be utilized to easily reproduce the corresponding results.

To evaluate the robustness and the feature selection capability, a noisy and high-dimensional synthetic dataset was considered. Specifically, 300 independent and identically distributed (i.i.d.) training samples with 300 i.i.d. testing samples were generated stochastically using the 500-dimensional standard multivariate Gaussian distribution (i.e., the covariance matrix is an identity matrix). The optimal solution was designed by a sparse format

$$\mathbf{w}^* = \overbrace{(w_1^*, \dots, w_{30}^*, 0, 0, 0, \dots, 0)}^{500 \text{ components}}^T \quad (25)$$

470 zeros

which means only the first 30 features are relevant to the task. The non-zero 30 elements in $\mathbf{w}^*$ was generated by the standard Gaussian distribution. The target $\mathbf{t}$ was obtained as the product between the covariates and the sparse solution $\mathbf{w}^*$. As a result, we trained the regression models on 300 training samples with 500 features, i.e., high-dimensional problem, and validated the models on the remaining 300 testing samples. For MCR-ARD, the kernel bandwidth h was selected by 5-fold cross validation from 30 values ranging from 1.0 to 1000 with equal differences on a logarithmic scale. Specifically, 300 training samples were divided into 5 sets. Each candidate value for kernel bandwidth was utilized for model training on 4 out of the 5 sets, and then evaluated on the other one, until each set is used for validation. The optimal kernel bandwidth was selected which realized the highest average correlation coefficient on 5 validation sets, and then used to train the model utilizing all training samples. Thus the testing data was not used in the kernel bandwidth selection.

Regarding the contamination for this toy dataset, we mainly considered the corrupted covariate matrix, because noisy brain recording is employed as input variable in brain decoding task. The conference paper [34] reported a simulation for corrupted target variable. To realize a realistic corruption on the covariate matrix $\mathbf{X}$, the arbitrary corruption model [43] was used, which means that any arbitrary elements in the covariate matrix could be contaminated randomly. To generate noises that are adverse enough on model training, the Laplace distribution was utilized for noise synthesis because the Laplace distribution is a heavy-tailed distribution. Hence, it is more likely to generate adverse outliers for data corruption. To be specific, a certain proportion (from 0 to 1.0 with a step 0.1) of the components for covariate matrix were randomly selected. Then, their values were added by a Laplace-distribution-induced noise. A zero-mean Laplace distribution with the following scale parameters was leveraged, respectively: 0.2, 0.3, 0.5, 0.7, 1.0, 1.2, and 1.5. A larger scale parameter is easier to induce the noise of large amplitude, thus being more detrimental to the dataset. On the other hand, only the training data was corrupted while the testing data remained unchanged, which is a typical strategy to assess the robustness for different algorithms [44]. To obtain generalized results, the above settings were executed for 100 Monte-Carlo repetitions.

2) Muscle Activity Reconstruction by Current Source:

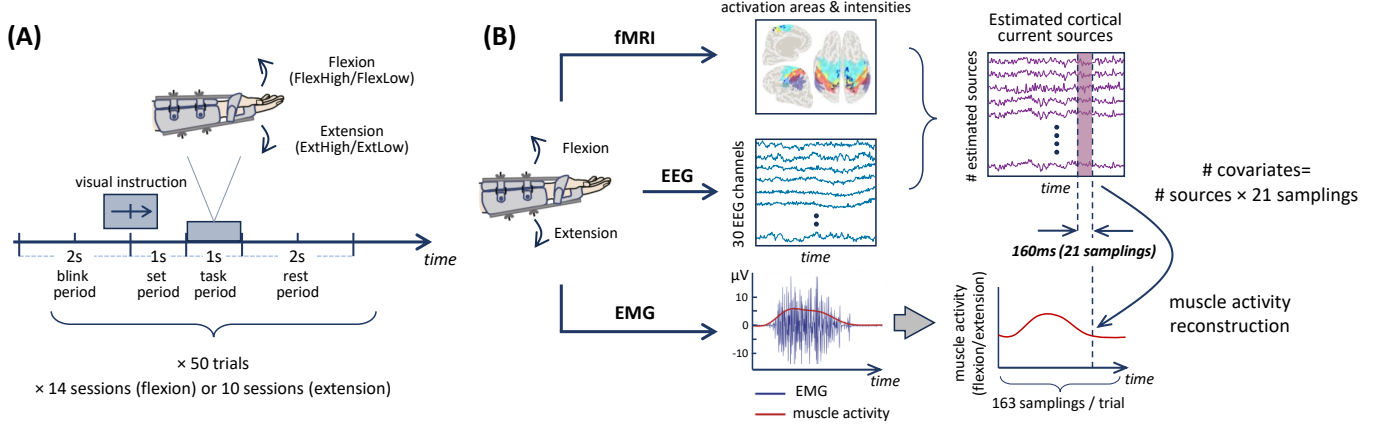
Five right-handed subjects participated in our previous study [16]. The experimental protocol was approved by a local ethics committee of the National Center of Neurology and Psychiatry, and all subjects gave written informed consent. Due to the data storage issue, only the data of four subjects was utilized in this study. Right wrist of each subject was immobilized with plastic

¹<https://sites.google.com/view/liyuanhao/code>

TABLE I: Summary of session information with the numbers of estimated sources for the muscle activity reconstruction dataset.

Task	Flexion														Extension											
Session	Sess.1	Sess.2	Sess.3	Sess.4	Sess.5	Sess.6	Sess.7	Sess.8	Sess.9	Sess.10	Sess.11	Sess.12	Sess.13	Sess.14	Sess.1	Sess.2	Sess.3	Sess.4	Sess.5	Sess.6	Sess.7	Sess.8	Sess.9	Sess.10		
Subject	Subj.1			Subj.2				Subj.3			Subj.4				Subj.1	Subj.2				Subj.3			Subj.4			
Activity	FH	FL	FL	FH	FL	FH	FL	FH	FL	FL	FH	FL	FH	FL	EL	EH	EL	EH	EL	EL	EH	EL	EH	EL	EL	
# Sources	63	63	59	77	77	130	79	137	137	117	211	211	199	159	59	127	127	130	79	117	112	112	199	159		

*FH=FlexHigh, FL=FlexLow, EH=ExtHigh, EL=ExtLow

**Fig. 2:** Schematic overview for the real-world muscle activity reconstruction task:(A) Trial diagram of the EEG-EMG experiment (B) Decoding paradigm of muscle activity reconstruction utilizing cortical current sources estimated from EEG and fMRI data.

casts that allowed the subject to perform isometric contraction tasks. The experiments were comprised of two parts, one using EEG and the other using fMRI. In addition, EMG signals were recorded in both two parts. During the EEG-EMG experiment, according to the visual instructions, each subject was requested to perform flexion and extension using high force or low force, thus resulting in the following four different combinations, i.e., 1) flexion with high force (FlexHigh), 2) flexion with low force (FlexLow), 3) extension with high force (ExtHigh), and finally 4) extension with low force (ExtLow). The muscle activity was recorded at the sampling rate of 5000 Hz by utilizing two pairs of 2 cm-spaced Ag/AgCl EMG electrodes attaching over right flexor carpi radialis and right extensor carpi radialis brevis, the major muscles for wrist flexion and extension. Simultaneously, a 30-channel EEG system positioned according to the extended international 10–20 system was utilized to record brain activity at the sampling rate of 5000 Hz. Two amplifiers for EMG and EEG recordings were both set to the resolution of $0.5 \mu V$ and the range of $\pm 16 mV$, respectively. Every trial for EEG-EMG experiment started with a 2 s blink period. Eye blink was only allowed in this period to avoid corruptions on EEG recordings. Then, the subjects received the visual instruction for 1 s, while the wrist kept still, called “set” period. In the next 1 s, subjects carried out the instructed action. Finally, the current trial ended by a 2 s rest, and the subsequent trial would start, as illustrated in Fig. 2(A). Each EEG-EMG session contained 50 trials with a same muscle activity. The session information is summarized in Table I for each subject with the activity category. Moreover, each subject participated in an fMRI experiment to identify the brain activation area and intensity of high spatial resolution for the cortical current source estimations. A 3 Tesla MRI scanner was used for the fMRI experiment, which consisted of 7 blocks for each subject. During each block, the subject was requested

to execute four muscle activities in a pseudo-randomized order.

For data preprocessing, the muscle activities were calculated from the EMG signals utilizing the method [45] for flexion and extension, which were further downsampled to 125 Hz for the following analysis. The last 0.3 s for the set period and all the task period were used for muscle activity reconstruction. Each trial had 163 samplings for the muscle activity as a result ($1.3 s \times 125 Hz$). The EEG recordings were first bandpass-filtered from 0.5 Hz to 40 Hz and then downsampled to 250 Hz. fMRI data was analyzed to create the area information that contained the regions of statistically significant voxels and identified the source estimation scope. Accordingly, the number of estimated current sources could be different. To estimate cortical current source, a hierarchical Bayesian method [46], [47] was utilized. The prior obtained from fMRI was used for the corresponding subject, i.e. individual analysis. Then, the estimated source was downsampled to 125 Hz which would be used as the covariate for the decoding. Table I lists the number of estimated sources.

To reconstruct the muscle activity from the brain recordings, a similar decoding paradigm as in our previous study [16] was utilized. First, an intra-session decoding was considered, where each session was used as a sub-dataset for both model training and testing. For intra-session decoding, only the corresponding muscle activities were reconstructed in performance validation, e.g., only the flexion was predicted for FlexHigh and FlexLow sessions. The target activity was normalized to a range of $[0, 1]$ for each session. Afterward each session was randomly divided into 5 batches, each containing 10 trials. The regression model was trained on 4 batches and then tested on the remaining one, until each batch was used for the testing. To predict the muscle activity for each time point, the cortical current sources during the previous 160 ms (21 samplings) were utilized as covariate, as shown in Fig. 2(B). To select a proper bandwidth for MCR-

TABLE II: Combination details of the inter-session framework.

Combo. Index	Subject	Task	Combined Sessions
Combo.1	Subj.1	FH vs. FL	Flexion Sess.1 + Flexion Sess.2
Combo.2		FL vs. EL	Flexion Sess.3 + Extension Sess.1
Combo.3	Subj.2	FH vs. FL	Flexion Sess.4 + Flexion Sess.5
Combo.4		EH vs. EL	Extension Sess.2 + Extension Sess.3
Combo.5		FH vs. EH	Flexion Sess.6 + Extension Sess.4
Combo.6	Subj.3	FL vs. EL	Flexion Sess.7 + Extension Sess.5
Combo.7		FH vs. FL	Flexion Sess.8 + Flexion Sess.9
Combo.8		FL vs. EL	Flexion Sess.10 + Extension Sess.6
Combo.9	Subj.4	FH vs. FL	Flexion Sess.11 + Flexion Sess.12
Combo.10		EH vs. EL	Extension Sess.7 + Extension Sess.8
Combo.11		FH vs. EH	Flexion Sess.13 + Extension Sess.9
Combo.12		FL vs. EL	Flexion Sess.14 + Extension Sess.10

*FH=FlexHigh, FL=FlexLow, EH=ExtHigh, EL=ExtLow

ARD, 35 random trials were selected from the training batches to train the model, with 30 different h values ranging from 1.0 to 1000, as mentioned in Section II-F.1. The optimal value was selected by validating on the remaining 5 trials for the training batches. Eventually, all 40 training trials were utilized to train MCR-ARD by the optimal kernel bandwidth. Thus, the tuning did not leverage the testing batch. As a result, for each session, five reconstruction models were trained by each algorithm, and 50 testing results were obtained by the 10 testing trials in each loop. Such a procedure was repeated 10 times to acquire more generalized results. Hence, each session resulted in 500 testing results for each algorithm.

Furthermore, for a more comprehensive decoding paradigm, inter-session decoding was considered to identify the decoding of different activities (flexion / extension) and different forces (high / low). Specifically, every two sessions of a same number of estimated sources were combined to train and test a muscle activity reconstruction model with a basically same framework as in the intra-session decoding. Table II lists the combinations of inter-session decoding. Notably after the training and testing batches were divided in each session, they were utilized jointly for each combination, that each had 80 training and 20 testing trials. For the decoding of different forces (Combos.1, 3, 4, 7, 9, 10), only the activated muscle activities were analyzed. For the decoding of two activities (Combos.2, 5, 6, 8, 11, 12), all muscle activities were utilized for prediction. Muscle activities were also normalized into $[0, 1]$. By 5-fold cross validation and 10 repetitions, each combination had 1000 testing results. The kernel bandwidth of MCR-ARD was selected in a similar way. Each candidate value was trained on 70 training trials and then validated on the other 10 training trials, which did not involve the testing trials.

3) Muscle Activity Reconstruction by EEG:

Furthermore, we evaluated the decoding performance for the above-mentioned muscle activity regression task, only utilizing the EEG signal in the decoding by LSR-ARD and MCR-ARD. Specifically, the fMRI recording and the cortical current source estimation procedure were excluded from the above-mentioned decoding paradigm. The pre-processed EEG recordings, which were also downsampled to 125 Hz similarly, were directly used as the covariates to predict the muscle activities, with the same intra-session decoding procedures, as are mentioned in Section II-F.2.

III. RESULTS

A. Synthetic Dataset

Fig. 3 illustrates the regression performance of the synthetic dataset obtained from LSR-ARD and MCR-ARD, respectively, averaged by 100 Monte-Carlo repetitions. The top row presents the correlation coefficient, while the second row shows the root mean squared error. One perceives that, the conventional LSR-ARD was influenced remarkably by the manual contamination. In contrast, MCR-ARD suppressed the adverse impact, leading to obviously higher correlation coefficient and lower root mean squared error, indicating a superior prediction than LSR-ARD.

Moreover, the feature selection capability was also evaluated for the synthetic toy. The third row in Fig. 3 shows the numbers of selected features for different corruption conditions. Further, to quantitatively examine the feature selection result, the recall was calculated for the 30 relevant dimensions, which indicates what percentage of these relevant features were selected by the sparse algorithm. The bottom row in Fig. 3 presents the feature selection recall for each algorithm. One can observe that LSR-ARD and MCR-ARD selected a similar number of covariates, and realized high enough recall when the noise proportion was 0. However, even a small noise proportion, such as 0.1, would result in an obvious impact on the number of selected features for LSR-ARD, whereas MCR-ARD could realize a much more stable number for selected features. Quantitatively, MCR-ARD achieved a higher feature selection recall than LSR-ARD under artificial corruption, suggesting better feature selection quality.

B. Muscle Activity Reconstruction by Current Source

Fig. 4 shows an example for intra-session decoding, Sess.14 for flexion activity reconstruction. Qualitatively, one perceives that the reconstruction from MCR-ARD is closer to the ground truth of the normalized muscle activity with a smaller standard deviation, in particular for the samplings from #120 to #163. For the quantitative performance indicators for all 24 sessions, the reconstruction result of intra-session decoding is illustrated in Fig. 5, with paired t -test to examine the statistical difference between LSR-ARD and MCR-ARD. Out of 14 flexion sessions the proposed MCR-ARD showed higher correlation coefficient and lower root mean squared error than the conventional LSR-ARD with statistically significant difference in 12 sessions. In the 10 extension sessions MCR-ARD outperformed LSR-ARD with significant difference for 8 and 9 sessions, for correlation coefficient and root mean squared error, respectively. Table III lists the average decoding performance for each type of muscle activity for each subject, where the superior results are marked in bold font. One observes that MCR-ARD outperformed LSR-ARD regarding all muscle activities on average for all subjects.

On the other hand, the spatio-temporal feature selection was studied for intra-session decoding. For each session, 50 models were trained (5-fold cross validation $\times$ 10 repetitions). In Fig. 6, as one would perceive, we illustrated the number of selected features averaged across 50 regression models for each session. LSR-ARD and MCR-ARD utilized a stable number of features within each session, while the numbers are obviously different between two algorithms. Concerning the temporal pattern, Fig. 7 shows the percentage that each temporal lag was leveraged in

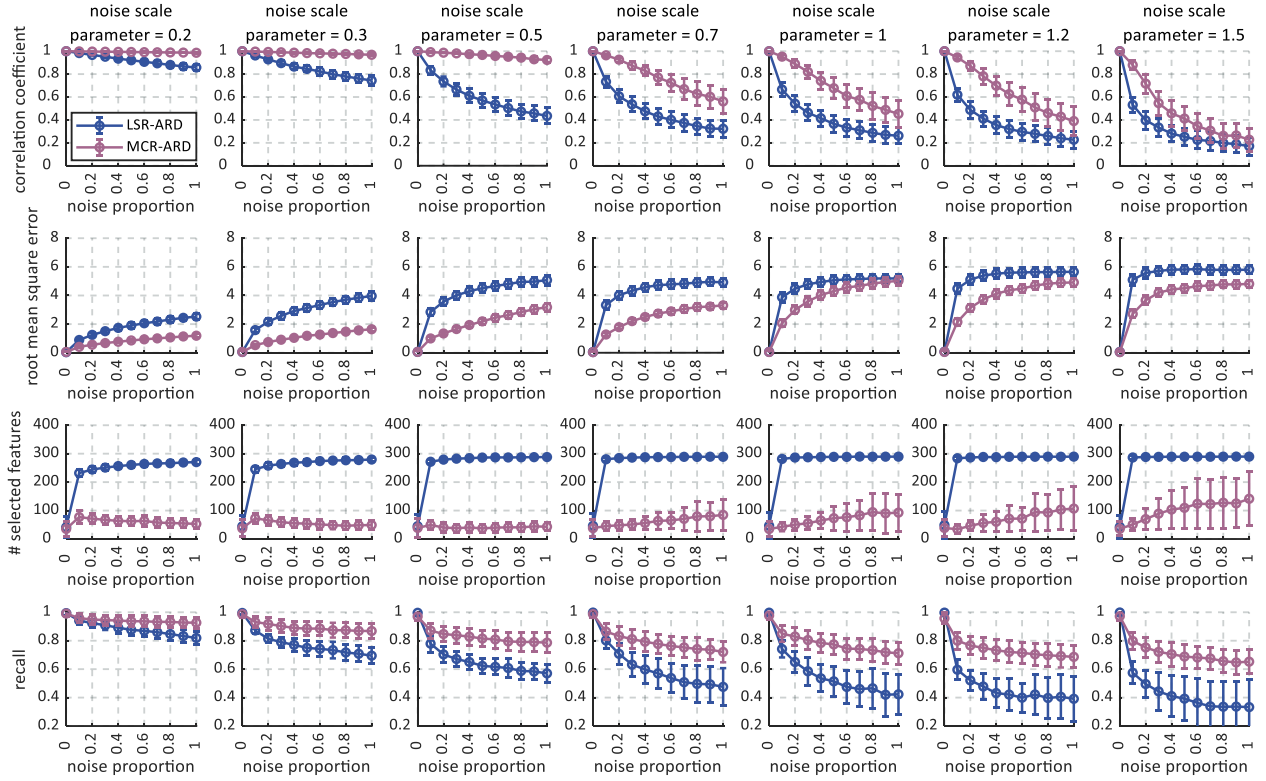


Fig. 3: Correlation coefficient, root mean squared error, number of selected features, and feature selection recall for the synthetic dataset by LSR-ARD and MCR-ARD, respectively. The results are averaged by 100 Monte-Carlo repetitions. Error bar presents the standard deviation for each result.

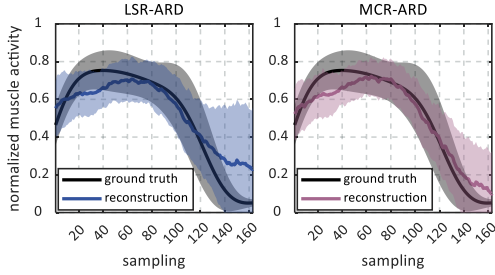


Fig. 4: Muscle activity of ground truth and reconstruction from LSR-ARD and MCR-ARD, respectively, for Sess.14 of flexion activity reconstruction. The ground truth illustrates the average of 50 trials in this session, and the reconstructions are averaged across 500 testing results. The shade area denotes the standard deviation.

50 models, averaged by all sessions of flexion and extension, respectively. For example, 50% means a temporal lag was used in 25 models. One can observe that, both LSR-ARD and MCR-ARD leveraged the whole 160 ms interval, whereas LSR-ARD had large fluctuations between neighboring lags. On the other hand, regarding spatial feature, the contribution of each source for predicting muscle activity was computed by

$$Contr(s) = \frac{\sum_t \|w_{s,t}\|}{\sum_s \sum_t \|w_{s,t}\|} \quad (26)$$

where $Contr(s)$ indicates the contribution of s -th source. $w_{s,t}$

corresponds to s -th source and t -th temporal lag. To test which algorithm realized superior source selection, the top 10 sources of the largest contributions for each algorithm were utilized for muscle activity reconstruction using the naive LSR (non-robust and non-sparse). Fig. 8 illustrates the result for each session. One can perceive that, even using the same LSR for prediction, the top 10 sources for MCR-ARD achieved higher correlation coefficient and lower root mean squared error with significant difference in most sessions for both flexion and extension. This suggests that, MCR-ARD can select those sources that contain more relevant information for brain decoding than LSR-ARD.

For the inter-session decoding, three examples of identifying different activities (flexion / extension) and forces (high / low) are illustrated in Fig. 9 for the proposed MCR-ARD algorithm. From Fig. 9(A)(B), one could perceive the obvious compliance of the reconstructed muscle activity for different force for both flexion and extension. Therefore, MCR-ARD could reconstruct muscle activity with a desired force in real-world applications. Moreover, Fig. 9(C) shows that MCR-ARD can reconstruct the activated and non-activated muscle activity for both flexion and extension simultaneously. This suggests that MCR-ARD could reconstruct the muscle activity for a more comprehensive task. Regarding the quantitative comparison between LSR-ARD and MCR-ARD, Fig.10 illustrates the inter-session decoding result for each combination. One observes that, except the correlation coefficients of flexion for Combo.3 and Combo.6, the proposed MCR-ARD outperformed the conventional LSR-ARD in every condition with statistically significant difference.

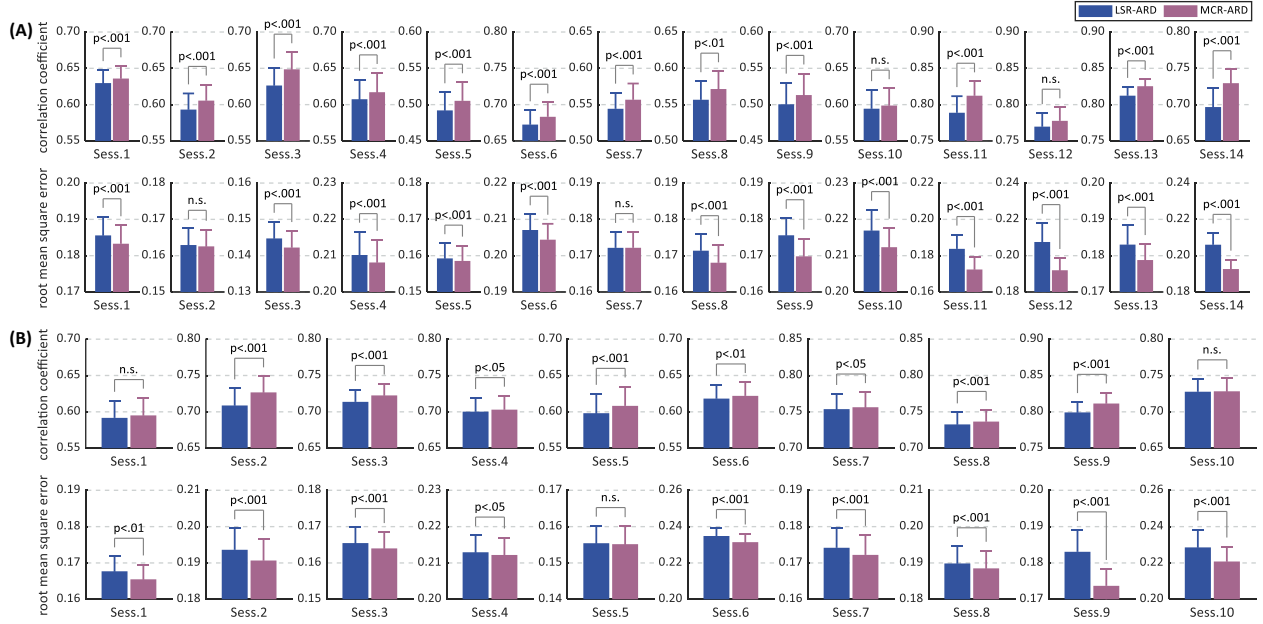


Fig. 5: Correlation coefficient and root mean squared error for the real-world muscle activity reconstruction dataset by employing the estimated cortical current source signals as the covariates: (A) flexion sessions (B) extension sessions. The results are averaged by repeating the five-fold cross validation for ten repetitions. The error bars denote the 95% confidence intervals for each result. Paired t -test was conducted to examine the statistically significant difference. “n.s.” indicates no significant difference ($p > .05$).

TABLE III: Average decoding performance regressing each muscle task of each subject. The superior results are marked in bold.

Performance indicator	Algorithm	Subj.1				Subj.2				Subj.3			Subj.4			
		FH	FL	EL		FH	FL	EH	EL	FH	FL	EL	FH	FL	EH	EL
Correlation coefficient $\uparrow$	LSR-ARD	0.6297	0.6098	0.5917		0.6400	0.5182	0.7045	0.6559	0.5567	0.5475	0.6182	0.8006	0.7332	0.7765	0.7302
	MCR-ARD	0.6359	0.6271	0.5951		0.6502	0.5310	0.7149	0.6654	0.5715	0.5560	0.6220	0.8188	0.7536	0.7839	0.7324
Root mean squared error $\downarrow$	LSR-ARD	0.1856	0.1538	0.1677		0.2086	0.1657	0.2033	0.1605	0.1714	0.1963	0.2348	0.1834	0.2068	0.1786	0.2092
	MCR-ARD	0.1833	0.1524	0.1655		0.2063	0.1654	0.2014	0.1596	0.1681	0.1911	0.2315	0.1756	0.1923	0.1730	0.2047

*FH=FlexHigh, FL=FlexLow, EH=ExtHigh, EL=ExtLow

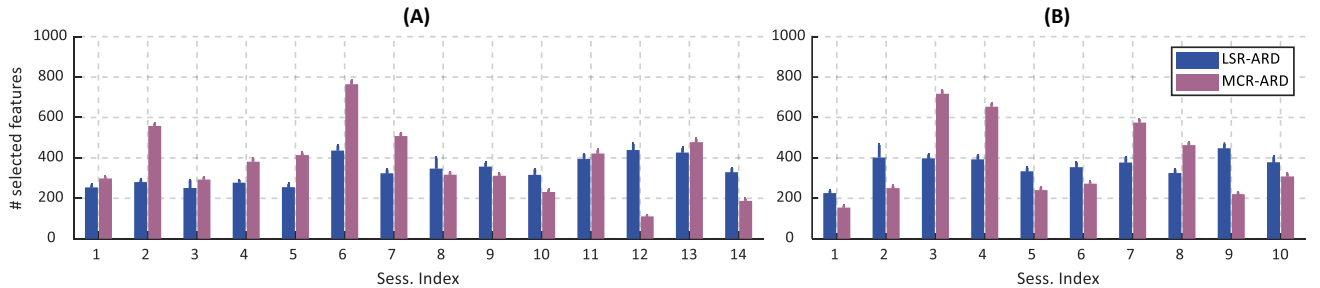


Fig. 6: Number of selected features averaged across the trained 50 reconstruction models for each session: (A) flexion sessions (B) extension sessions. The error bars indicate the standard deviations for each corresponding result. Note that the total number of features would be different across sessions because the number of estimated current sources varies, as summarized in Table-I.

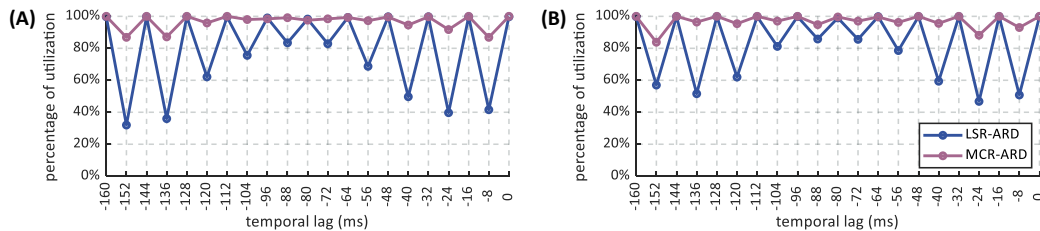


Fig. 7: Percentage of utilization for each temporal lag in the trained 50 reconstruction models averaged across all sessions: (A) flexion sessions (B) extension sessions.

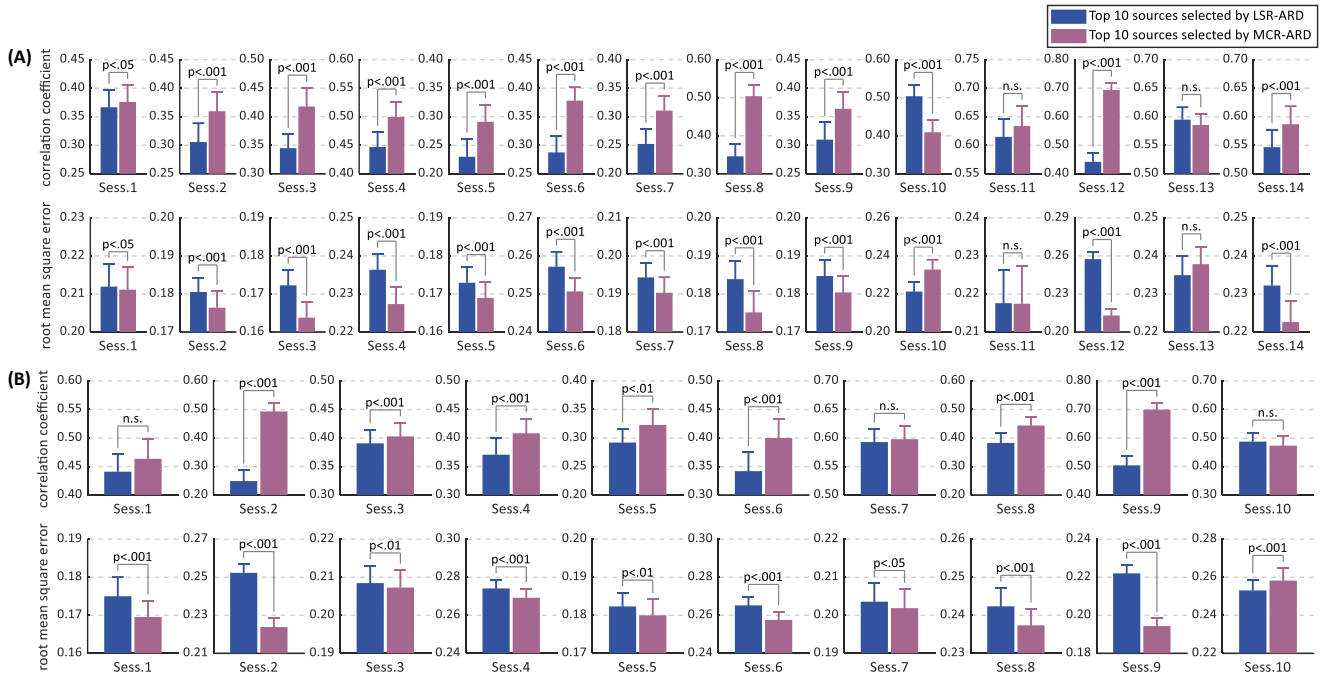


Fig. 8: Correlation coefficient and root mean squared error for the real-world muscle activity reconstruction dataset employing the top 10 current sources with the largest contributions as the covariates and using the naive LSR algorithm for reconstruction: (A) flexion sessions (B) extension sessions. The results are averaged by repeating the five-fold cross validation for ten repetitions. The error bars denote the 95% confidence intervals for each result. Paired t -test was used to examine the statistically significant difference. “n.s.” indicates no significant difference ($p > .05$). The top 10 current sources were selected for each session separately.

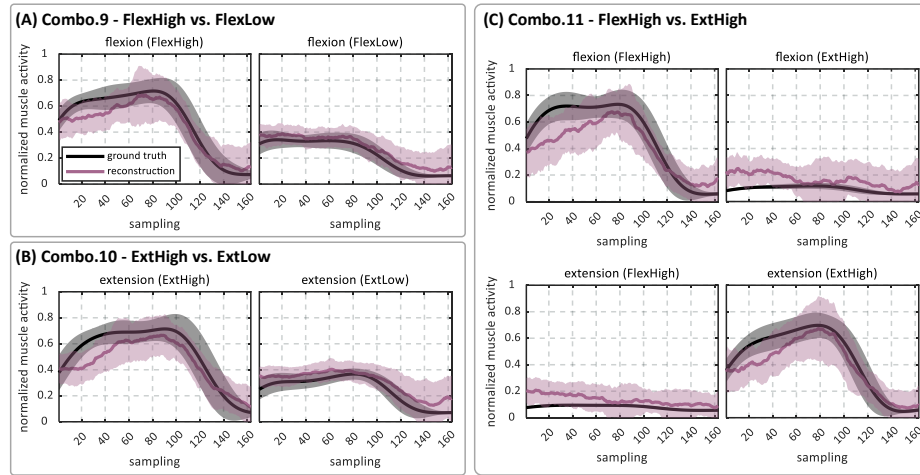


Fig. 9: Three examples of inter-session decoding by MCR-ARD: (A) Combo.9 (B) Combo.10 (C) Combo.11. The ground truth illustrates the average muscle activity of 50 trials for the corresponding session, while the reconstructions are averaged across 500 testing results of each session (totally 1000 testing results in each combo). The shade area denotes the standard deviation.

C. Muscle Activity Reconstruction by EEG

In EEG-based intra-session decoding, Fig. 11 illustrates the reconstruction result for each session. One can observe, similar to before, that the proposed MCR-ARD achieved significantly superior prediction than LSR-ARD on the majority of sessions. Specifically, for 14 flexion sessions, MCR-ARD outperformed LSR-ARD with significant difference for all sessions regarding correlation coefficient, and for 11 sessions regarding root mean squared error. For 10 extension sessions, MCR-ARD achieved

significantly higher correlation coefficient and lower root mean squared error for all sessions and 9 sessions, respectively. This further authenticates the admirable competence of MCR-ARD for decoding sophisticated activity using noisy brain recording.

IV. DISCUSSION

This study proposed a novel robust sparse regression method named as MCR-ARD, which provides a powerful tool for brain activity decoding. Although the conference paper for this study

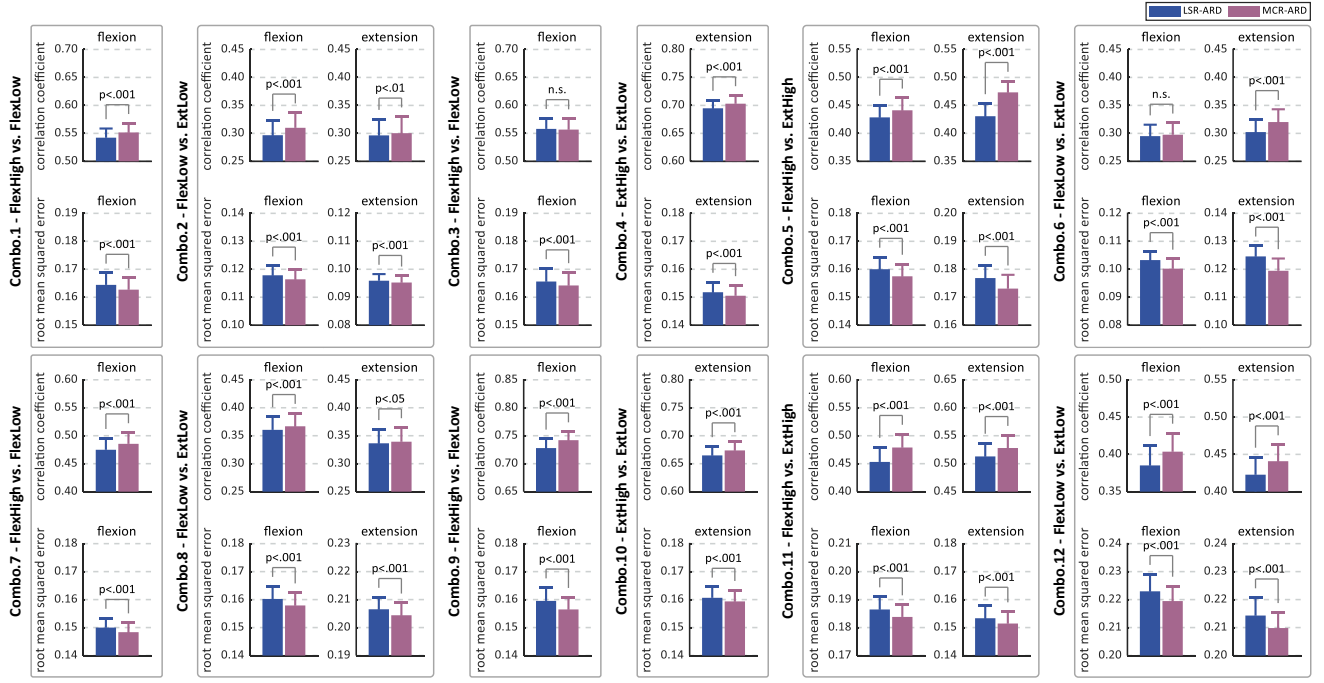


Fig. 10: Correlation coefficient and root mean squared error for the real-world muscle activity reconstruction dataset by employing the estimated cortical current source signals as the covariates under the inter-session decoding framework. The error bars denote the 95% confidence intervals. Paired t -test was used to examine the statistical difference. “n.s.” indicates no significant difference ($p > .05$).

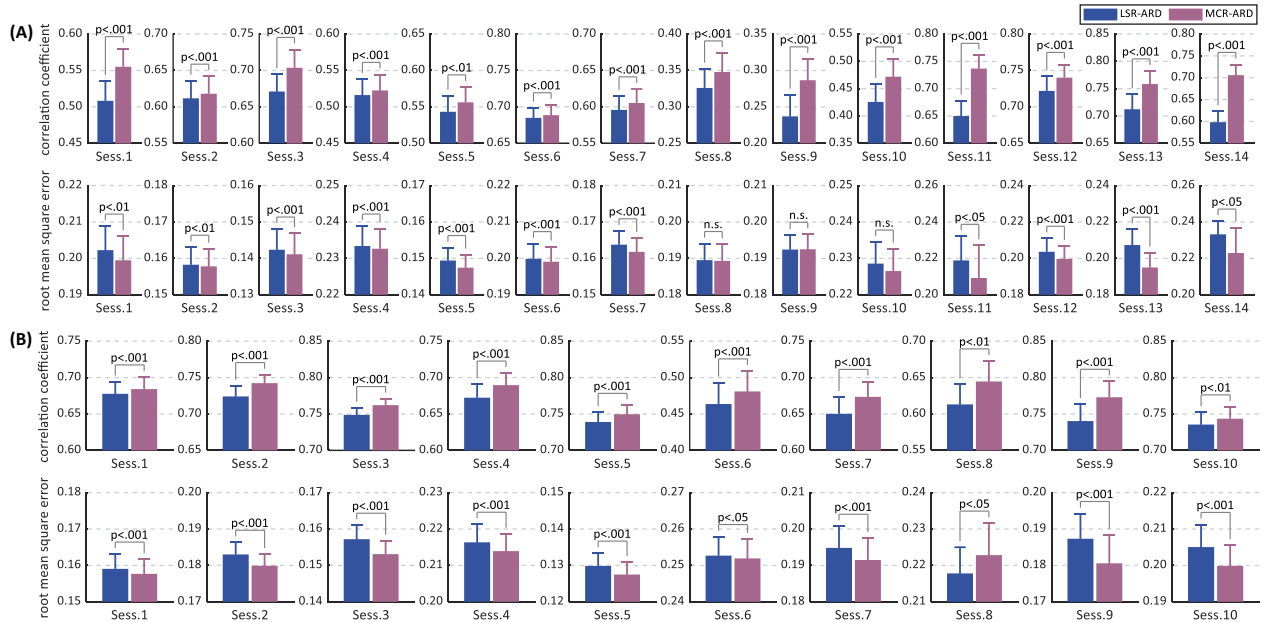


Fig. 11: Correlation coefficient and root mean squared error for the real-world muscle activity reconstruction dataset by employing the preprocessed 30-channel EEG signals as the covariates: (A) flexion sessions (B) extension sessions. The results are averaged by repeating the five-fold cross validation for ten repetitions. The error bars denote the 95% confidence intervals for each result. Paired t -test was conducted to examine the statistically significant difference. “n.s.” indicates no significant difference ($p > .05$).

[34] has introduced the theoretical derivation for the algorithm and used a synthetic example with noisy output for evaluation, it remained unclear whether MCR-ARD could realize superior

neural decoding performance in a real-world application. Since the noisy brain recording is used as the covariate for decoding, the present paper first employed a synthetic dataset with noisy

covariate, which is a more realistic problem setting for the real-world brain activity decoding. To evaluate the proposed MCR-ARD algorithm with real brain datasets, this paper further used the muscle activity reconstruction scenario with current source and EEG signal. The experimental results in this paper suggest that, our proposed MCR-ARD algorithm could provide a better substitute for brain signal decoding instead of the conventional sparse Bayesian learning method. Furthermore, because MCR-ARD was proposed in a primary perspective, i.e. the likelihood function, it can be conveniently employed with other Bayesian techniques to realize further improvement. For example, block-sparse Bayesian learning (BSBL) has been employed for EEG source estimation [22], since block-sparsity could better model the EEG signals which are more likely to generate from locally synchronized neural issues. Integrating the proposed likelihood function (12) with BSBL is expected to be more adequate since the proposed likelihood function will help solve the intractable problem of EEG recording noises. The proposed robust sparse regression algorithm is in particular applicable for the scenario where it is difficult to collect a large amount of brain data, e.g., collecting both brain and muscle data meanwhile in the muscle activity reconstruction task. Complicated neural network, such as the convolutional neural network (CNN) and the long short-term memory (LSTM), will be difficult to train for such a case, while our proposed method could provide a practical approach. MCR-ARD could be utilized to structure a more advanced BCI system for continuous motion prediction. For example, muscle activity can be predicted from brain recordings by MCR-ARD, which could be further leveraged to control the prosthetic hand in real time for motion assistance [8], [9]. MCR-ARD will help improve the BCI systems to realize better motion reproduction and rehabilitation.

Robustness and sparseness are two crucial natures for many machine learning tasks. Existing robust sparse methods mainly focus on regularizing a robust objective function by a sparsity-promoting regularization term, e.g., L_1 -regularized MCC [48], [49]. However, these methods have two hyper-parameters, one controlling robustness and another one controlling sparseness. For example, L_1 -regularized MCC tunes kernel bandwidth for MCC and regularization weight for L_1 -norm. This usually asks for a grid search to find the optimal configurations for the two hyper-parameters, which can be considerably time-consuming in practice. To reduce hyper-parameter, we leveraged the ARD method which can realize adaptive sparseness. However, ARD does not have a factorable regularization form [42]. Hence, we cannot combine MCC and ARD directly like existing methods. To address this obstacle, we innovatively exposed the inherent error assumption in MCC with an explicit form $\mathcal{C}(\epsilon|0, h)$. Thus we can integrate robust MCC with ARD in a Bayesian regime. To the best of our knowledge, this was the first time to explain MCC from a Bayesian viewpoint. Compared to L_1 -regularized MCR, our proposed MCR-ARD can realize better performance for the noisy and high-dimensional regression, as shown in the preliminary conference paper [34]. More crucially, MCR-ARD only has one hyperparameter to tune, i.e. the kernel bandwidth, which will greatly facilitate its applications to real-world tasks. The reason why we did not assess L_1 -regularized MCR in this paper was it needed unacceptable time to finish the grid search

to select two hyper-parameters. For example, if we select both kernel bandwidth and regularization weight from 30 candidate values, we need to train the model with 900 combinations. By comparison, MCR-ARD only needs a line search to determine h from 30 candidates. As a result, MCR-ARD greatly reduces the time to select the hyperparameter, which makes it evidently easier to use in the real-world brain activity decoding scenario.

The core insight of this research is that our proposed deviant error assumption $\mathcal{C}(\epsilon|0, h)$ realizes much better robustness than the conventional Gaussian distribution regarding the utilization for the likelihood function. The decisive property of $\mathcal{C}(\epsilon|0, h)$, as mentioned in Section II-D, is $\lim_{\epsilon \rightarrow \infty} \mathcal{C}(\epsilon|0, h) = 1$, while the generic distributions will converge to 0. We desire to argue that, it is this difference that greatly ameliorates the robustness. To be specific, if a dataset contains adverse noises, they would result in training errors with arbitrarily large amplitudes. Thus, the probability of a large error occurring is not zero, conflicting with the traditional error assumptions. In contrast, our deviant $\mathcal{C}(\epsilon|0, h)$ assumes that any arbitrarily large error might happen with a non-zero probability, which is more sensible for a noisy scenario. On the other hand, by tradition, improper distribution is only allowed for the prior distribution for Bayesian learning [35]. This study employed the deviant assumption $\mathcal{C}(\epsilon|0, h)$ for the likelihood, and thus realized significant improvements. The integration of MCC into the Bayesian learning framework will be further scrutinized in our future works, with the motivations of recent studies that have made effort to establish a theoretical guarantee for the non-normalized likelihood models [50], [51].

We would also like to discuss about the only hyperparameter in our MCR-ARD, the kernel bandwidth h . It was determined by utilizing cross validation (Section II-F.1) or an independent validation set (Section II-F.2 and II-F.3). As mentioned before, h was regarded as a fixed parameter in our algorithm. It would be an interesting question whether we can make h as a random variable and incorporate it in the Bayesian inference processes, so that h could be determined automatically to realize adaptive robustness. This topic will be investigated in our future studies.

In addition, we also investigated how well MCR-ARD could be generalized for datasets with different scales. We leveraged the synthetic dataset, as mentioned in Section II-F.1, and used different numbers of samples and different dimensions to study how the training time was affected. Fig. 12 shows the average computation time for one iteration of free energy minimization (step 4-step 10 in Algorithm 1) obtained by MATLAB R2023b (MathWorks Inc.) on Intel Core i9-13900H CPU, 32GB RAM, and Windows 11 OS. One could find that the computation time increased linearly with respect to the number of samples, since both w -step (21) and the calculation of negative Hessian matrix (19) have linear complexity in regard to N . On the other hand, MCR-ARD has a polynomial complexity regarding dimension D , resulting from w -step (21) of $\mathcal{O}(ND^2 + D^3)$, and Hessian matrix inverse of $\mathcal{O}(D^3)$. This indicates that MCR-ARD could be well applied for various scenarios with different data scales.

Finally, for the proposed MCR-ARD algorithm in this study, the limitations are discussed as follows. The critical difference between MCR-ARD and conventional LSR-ARD is that, LSR-ARD utilizes the Gaussian error assumption that may be overly idealistic whereas MCR-ARD employs the deviant assumption

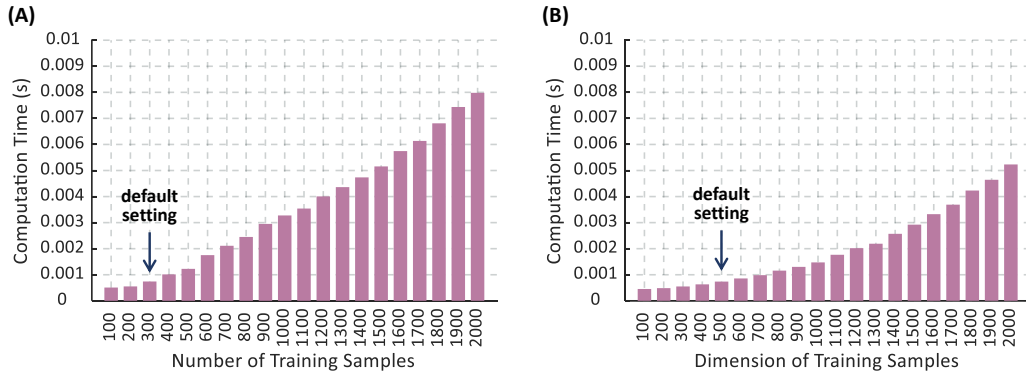


Fig. 12: Average computation time for one free energy minimization iteration under different dataset scales: (A) under different numbers of training samples (the dimension is fixed to be 500). (B) under different dimensions of training samples (the number of training samples is fixed to be 300). The number of the relevant features is fixed to be 30 as in the original synthetic example.

$\mathcal{C}(\epsilon|0, h)$ in the likelihood function with heavier tail. However, it will be hard to guarantee that $\mathcal{C}(\epsilon|0, h)$ is always the optimal choice for the error assumption in every context. For example, from Fig. 1, one can observe that $\mathcal{C}(\epsilon|0, h)$ cannot approximate the Gaussian distribution naturally. As a result, $\mathcal{C}(\epsilon|0, h)$ would be less adequate and lead to a biased solution for the problems with a Gaussian error distribution. On the other hand, $\mathcal{C}(\epsilon|0, h)$ with one single peak is not appropriate for a more complicated error distribution, e.g., the multi-peak error distribution. Using a multi-peak model may further help improve the brain activity decoding accuracy [52], [53]. Hence, it would be an interesting future work to investigate an improved error assumption which has a better flexibility while maintains the desirable robustness.

V. CONCLUSION

To realize robust sparse regression, this research first derived an explicit error assumption $\mathcal{C}(\epsilon|0, h)$ inherent in robust MCR, and utilized it to construct a robust likelihood function. Using ARD technique, MCR was integrated with the sparse Bayesian learning framework. MCR-ARD was systematically evaluated by a synthetic dataset with noisy input and a real-world muscle activity reconstruction task with two different brain modalities. The experimental results show that our proposed algorithm can largely ameliorate the robustness for sparse Bayesian learning. The proposed **sparse Bayesian correntropy learning** method, including our previous work [32], provides a powerful tool for the real-world brain activity decoding tasks to realize advanced brain-computer interface technology.

REFERENCES

- [1] U. Chaudhary, N. Birbaumer, and A. Ramos-Murguialday, "Brain-computer interfaces for communication and rehabilitation," *Nature Reviews Neurology*, vol. 12, no. 9, pp. 513–525, 2016.
- [2] A. Kawala-Sterniuk, N. Browarska, A. Al-Bakri, M. Pelc, J. Zygarlicki, M. Sidikova, R. Martinek, and E. J. Gorzelanczyk, "Summary of over fifty years with brain-computer interfaces—a review," *Brain Sciences*, vol. 11, no. 1, p. 43, 2021.
- [3] M. A. Cervera, S. R. Soekadar, J. Ushiba, J. d. R. Millán, M. Liu, N. Birbaumer, and G. Garipelli, "Brain-computer interfaces for post-stroke motor rehabilitation: a meta-analysis," *Annals of Clinical and Translational Neurology*, vol. 5, no. 5, pp. 651–663, 2018.
- [4] P. Wang, X. Cao, Y. Zhou, and D. Zhang, "A comprehensive review on motion trajectory reconstruction for eeg-based brain-computer interface," *Frontiers in Neuroscience*, vol. 17, p. 1086472, 2023.
- [5] P. Arpaia, A. Esposito, A. Natalizio, and M. Parvis, "How to successfully classify eeg in motor imagery bci: a metrological analysis of the state of the art," *Journal of Neural Engineering*, vol. 19, no. 3, p. 031002, 2022.
- [6] L. Bi, C. Guan *et al.*, "A review on emg-based motor intention prediction of continuous human upper limb motion for human-robot collaboration," *Biomedical Signal Processing and Control*, vol. 51, pp. 113–127, 2019.
- [7] Y. Koike, "Motion estimation from surface emg signals using multi-array electrodes," in *Biomedical Engineering*. Jenny Stanford Publishing, 2024, pp. 133–151.
- [8] Z. Qin, S. Stapornchaisit, Z. He, N. Yoshimura, and Y. Koike, "Multi-joint angles estimation of forearm motion using a regression model," *Frontiers in Neuroinformatics*, vol. 15, p. 685961, 2021.
- [9] Z. Qin, Z. He, Y. Li, S. Saetia, and Y. Koike, "A cw-cnn regression model-based real-time system for virtual hand control," *Frontiers in Neuroinformatics*, vol. 16, p. 1072365, 2022.
- [10] W.-k. Tam, T. Wu, Q. Zhao, E. Keefer, and Z. Yang, "Human motor decoding from neural signals: a review," *BMC Biomedical Engineering*, vol. 1, pp. 1–22, 2019.
- [11] T. Umeda, M. Koizumi, Y. Katakai, R. Saito, and K. Seki, "Decoding of muscle activity from the sensorimotor cortex in freely behaving monkeys," *NeuroImage*, vol. 197, pp. 512–526, 2019.
- [12] J. Wang, L. Bi, and W. Fei, "Eeg-based motor bcis for upper limb movement: Current techniques and future insights," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 4413–4427, 2023.
- [13] T. Das, L. Gohain, N. M. Kakoty, M. Malarvili, P. Widiyanti, and G. Kumar, "Hierarchical approach for fusion of electroencephalography and electromyography for predicting finger movements and kinematics using deep learning," *Neurocomputing*, vol. 527, pp. 184–195, 2023.
- [14] G. Ganesh, E. Burdet, M. Haruno, and M. Kawato, "Sparse linear regression for reconstructing muscle activity from human cortical fmri," *NeuroImage*, vol. 42, no. 4, pp. 1463–1472, 2008.
- [15] A. Toda, H. Imamizu, M. Kawato, and M.-a. Sato, "Reconstruction of two-dimensional movement trajectories from selected magnetoencephalography cortical currents by combined sparse bayesian methods," *NeuroImage*, vol. 54, no. 2, pp. 892–905, 2011.
- [16] N. Yoshimura, C. S. DaSalla, T. Hanakawa, M.-a. Sato, and Y. Koike, "Reconstruction of flexor and extensor muscle activities from electroencephalography cortical currents," *NeuroImage*, vol. 59, no. 2, pp. 1324–1337, 2012.
- [17] N. Yoshimura, H. Tsuda, T. Kawase, H. Kambara, and Y. Koike, "Decoding finger movement in humans using synergy of eeg cortical current signals," *Scientific Reports*, vol. 7, p. 11382, 2017.
- [18] A. Mejia Tobar, R. Hyoudou, K. Kita, T. Nakamura, H. Kambara, Y. Ogata, T. Hanakawa, Y. Koike, and N. Yoshimura, "Decoding of ankle flexion and extension from cortical current sources estimated from non-invasive brain activity recording methods," *Frontiers in Neuroscience*, vol. 11, p. 733, 2018.
- [19] R. Sosnik and L. Zheng, "Reconstruction of hand, elbow and shoulder actual and imagined trajectories in 3d space using eeg current source dipoles," *Journal of Neural Engineering*, vol. 18, p. 056011, 2021.
- [20] H. Altafheri, G. Muhammad, M. Alsulaiman, S. U. Amin, G. A. Altuwai-

- jri, W. Abdul, M. A. Bencherif, and M. Faisal, "Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review," *Neural Computing and Applications*, vol. 35, no. 20, pp. 14 681–14 722, 2023.
- [21] A. Hashemi, C. Cai, G. Kutyniok, K.-R. Müller, S. S. Nagarajan, and S. Haufe, "Unification of sparse bayesian learning algorithms for electromagnetic brain imaging with the majorization minimization framework," *NeuroImage*, vol. 239, p. 118309, 2021.
- [22] M. Qu, C. Chang, J. Wang, J. Hu, and N. Hu, "Nonnegative block-sparse bayesian learning algorithm for eeg brain source localization," *Biomedical Signal Processing and Control*, vol. 77, p. 103838, 2022.
- [23] W. Wang, F. Qi, D. P. Wipf, C. Cai, T. Yu, Y. Li, Y. Zhang, Z. Yu, and W. Wu, "Sparse bayesian learning for end-to-end eeg decoding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 15 632–15 649, 2023.
- [24] T. Ball, M. Kern, I. Mutschler, A. Aertsen, and A. Schulze-Bonhage, "Signal quality of simultaneously recorded invasive and non-invasive eeg," *NeuroImage*, vol. 46, no. 3, pp. 708–716, 2009.
- [25] T. T. Liu, "Noise contributions to the fmri signal: An overview," *NeuroImage*, vol. 143, pp. 141–151, 2016.
- [26] T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M. J. McKeown, V. Iragui, and T. J. Sejnowski, "Removing electroencephalographic artifacts by blind source separation," *Psychophysiology*, vol. 37, no. 2, pp. 163–178, 2000.
- [27] J. C. Principe, *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.
- [28] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-gaussian signal processing," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5286–5298, 2007.
- [29] Y. Li, B. Chen, G. Wang, N. Yoshimura, and Y. Koike, "Partial maximum correntropy regression for robust electrocorticography decoding," *Frontiers in Neuroscience*, vol. 17, p. 1213035, 2023.
- [30] A. Singh, R. Pokharel, and J. Principe, "The c-loss function for pattern classification," *Pattern Recognition*, vol. 47, no. 1, pp. 441–453, 2014.
- [31] G. Xu, B.-G. Hu, and J. C. Principe, "Robust c-loss kernel classifiers," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 3, pp. 510–522, 2016.
- [32] Y. Li, B. Chen, Y. Shi, N. Yoshimura, and Y. Koike, "Correntropy-based logistic regression with automatic relevance determination for robust sparse brain activity decoding," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 8, pp. 2416–2429, 2023.
- [33] R. He, B.-G. Hu, W.-S. Zheng, and X.-W. Kong, "Robust principal component analysis based on maximum correntropy criterion," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1485–1494, 2011.
- [34] Y. Li, B. Chen, O. Yamashita, N. Yoshimura, and Y. Koike, "Adaptive sparseness for correntropy-based robust regression via automatic relevance determination," in *2023 International Joint Conference on Neural Networks (IJCNN)*, 2023, pp. 1–8.
- [35] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*. Chapman and Hall/CRC, 1995.
- [36] M. Tipping, "The relevance vector machine," in *Advances in Neural Information Processing Systems*, vol. 12, 1999.
- [37] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.
- [38] M. A. Figueiredo, "Adaptive sparseness for supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1150–1159, 2003.
- [39] C. M. Bishop and M. E. Tipping, "Variational relevance vector machines," in *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 2000, pp. 46–53.
- [40] I. Santamaría, P. P. Pokharel, and J. C. Principe, "Generalized correlation function: definition, properties, and application to blind equalization," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2187–2197, 2006.
- [41] B. Chen, J. Wang, H. Zhao, N. Zheng, and J. C. Principe, "Convergence of a fixed-point algorithm under maximum correntropy criterion," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1723–1727, 2015.
- [42] D. Wipf and S. Nagarajan, "A new view of automatic relevance determination," in *Advances in Neural Information Processing Systems*, vol. 20, 2007.
- [43] Y. Chen, C. Caramanis, and S. Mannor, "Robust sparse regression under adversarial corruption," in *Proceedings of the 30th International Conference on Machine Learning*, vol. 28, 2013, pp. 774–782.
- [44] X. Zhu and X. Wu, "Class noise vs. attribute noise: A quantitative study," *Artificial Intelligence Review*, vol. 22, pp. 177–210, 2004.
- [45] Y. Koike and M. Kawato, "Estimation of dynamic joint torques and trajectory formation from surface electromyography signals using a neural network model," *Biological Cybernetics*, vol. 73, no. 4, pp. 291–300, 1995.
- [46] M.-a. Sato, T. Yoshioka, S. Kajihara, K. Toyama, N. Goda, K. Doya, and M. Kawato, "Hierarchical bayesian estimation for meg inverse problem," *NeuroImage*, vol. 23, no. 3, pp. 806–826, 2004.
- [47] T. Yoshioka, K. Toyama, M. Kawato, O. Yamashita, S. Nishina, N. Yamagishi, and M.-a. Sato, "Evaluation of hierarchical bayesian method through retinotopic brain activities reconstruction from fmri and meg signals," *NeuroImage*, vol. 42, no. 4, pp. 1397–1413, 2008.
- [48] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1561–1576, 2010.
- [49] W. Ma, H. Qu, G. Gui, L. Xu, J. Zhao, and B. Chen, "Maximum correntropy criterion based sparse adaptive filtering algorithms for robust channel estimation under non-gaussian environments," *Journal of the Franklin Institute*, vol. 352, no. 7, pp. 2708–2727, 2015.
- [50] F. Giummolè, V. Mamelì, E. Ruli, and L. Ventura, "Objective bayesian inference with proper scoring rules," *Test*, vol. 28, no. 3, pp. 728–755, 2019.
- [51] T. Matsuda, M. Uehara, and A. Hyvärinen, "Information criteria for non-normalized models," *Journal of Machine Learning Research*, vol. 22, no. 158, pp. 1–33, 2021.
- [52] B. Chen, Y. Li, J. Dong, N. Lu, and J. Qin, "Common spatial patterns based on the quantized minimum error entropy criterion," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 11, pp. 4557–4568, 2020.
- [53] Y. Li, B. Chen, N. Yoshimura, and Y. Koike, "Restricted minimum error entropy criterion for robust classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 11, pp. 6599–6612, 2022.