Partial Maximum Correntropy Regression for Robust Trajectory Decoding from Noisy Epidural Electrocorticographic Signals

Yuanhao Li, Badong Chen, Senior Member, IEEE, Gang Wang, Member, IEEE, Natsue Yoshimura, and Yasuharu Koike

Abstract—The Partial Least Square Regression (PLSR) algorithm exhibits exceptional competence for predicting continuous variables from inter-correlated brain recordings in brain-computer interfaces, which achieved successful prediction from epidural electrocorticography of macaques to three-dimensional continuous hand trajectories recently. Nevertheless, PLSR is in essence formulated based on the least square criterion, thus, being non-robust with respect to complicated noises consequently. The aim of the present study is to propose a robust version of PLSR. To this end, the maximum correntropy criterion is adopted to structure a new robust variant of PLSR, namely Partial Maximum Correntropy Regression (PMCR). Half-quadratic optimization technique is utilized to calculate the robust latent variables. We assess the proposed PMCR on a synthetic example and the public Neurotycho dataset. Compared with the conventional PLSR and the state-of-the-art variant, PMCR realized superior prediction competence on three different performance indicators with contaminated training set. The proposed PMCR was demonstrated as an effective approach for robust decoding from noisy brain measurements, which could reduce the performance degradation resulting from adverse noises, thus, improving the decoding robustness of brain-computer interfaces.

Index Terms—brain-computer interface, epidural electrocorticography, partial least square regression, maximum correntropy criterion, robustness.

I. INTRODUCTION

Brain-computer interface (BCI) has been conceived as a promising technology that translates cerebral recordings generated by cortical neurons into appropriate commands for controlling neuroprosthetic devices [1]. The capability of BCI for repairing or reproducing sensory-motor functions has been increasingly intensified by recent scientific and technological advances [2]–[4]. The noninvasive recordings, especially electroencephalogram (EEG) and magnetoencephalogram (MEG), are widely exploited to structure BCI systems due to their ease of use and satisfactory temporal resolution, whereas the non-invasive BCI systems could be limited in their capabilities and customarily require considerable training [5]. Invasive single-unit activities and local field potentials usually provide better decoding performance, which suffer pessimistic long-term stability however due to capriciousness in the recorded neuronal ensembles [6]. A sophisticated alternative that exhibits higher signal amplitudes than EEG while presents superior long-term stability compared to invasive modalities, is the semi-invasive electrocorticography (ECoG) [7]. Many studies in recent years have investigated the potentials of ECoG recording in decoding discrete hand movements [8]–[10]. The serviceability of ECoG for online practice was also demonstrated in [9], [11], [12].

ECoG can be classified into two categories: subdural ECoG (SECoG), which records the neural electrical signals by placing the electrodes under the dura, and epidural ECoG (ECoG), which records signals on the dura. Although SECoG provides better quality of recordings, the surgical complications obstruct its extensive application, including intracerebral hemorrhage, subdural hematoma, cerebral edema, and infarction [13]–[15]. Consequently, ECoG has been capturing increasing attention in the BCI community. A recent study has realized successful prediction from the ECoG signals of two Japanese macaques during an asynchronous food-reaching task, to the continuous three-dimensional hand trajectories [16]. The utilized decoding method was the partial least square regression (PLSR), which was originally developed for econometrics and chemometrics [17], and has emerged as a prevailing method for neurological imaging and decoding [18].

PLSR algorithm answers the regression problems primarily for dimensional reduction techniques both on independent and dependent variables, in which the dimensionality-reduced samples (usually called as latent variables) of respective sets exhibit maximal correlation, thus structuring association from independent to dependent variables. Multi-way PLSR, NPLSR in short, was proposed as a generalization for tensor variables [19]. Recently, investigators have conducted further studies to ameliorate the performance of PLSR, most of which calculate
the prediction model with an additive regularization item [20], [21]. Nevertheless, PLSR essentially utilizes the least square criterion, which probably assigns large importance to deviated noises, as a result, leading to poor robustness.

In the present study, we aim to propose a new robust version of PLSR by introducing the maximum correntropy criterion (MCC) to replace the non-robust least square criterion, which was proposed in the information theoretic learning (ITL) [22], and has achieved state-of-the-art robust approaches in different tasks, including regression [23]–[25], classification [26], [27], principal component analysis [28], and feature extraction [29].

The remainder of this paper is organized as follows. Section II introduces the mathematical derivation of the conventional PLSR algorithm, with its existing improved version. In Section III we present a brief introduction about MCC, and reformulate PLSR with MCC. In Section IV we evaluate the proposed method on synthetic and real eECoG datasets, respectively. In Section V we conclude this paper.

II. PARTIAL LEAST SQUARE REGRESSION

A. Conventional PLSR

Consider the explanatory matrix (or independent variables) $X \in \mathbb{R}^{L \times N}$ and the response matrix (or dependent variables) $Y \in \mathbb{R}^{L \times M}$, in which $L$ denotes the number of observations with $N$ and $M$ being the respective dimensions. PLSR is an iterative algorithm, which executes dimensionality reduction simultaneously on both explanatory and response matrices, so that the acquired latent variables correlate maximally at each iteration. In the $s$-th iteration, where $X_s$ and $Y_s$ are the current residual matrices, PLSR calculates two projectors $w_s \in \mathbb{R}^N$ and $c_s \in \mathbb{R}^M$ to acquire the latent variables $t_s = X_sw_s$ and $u_s = Y_sc_s$ by maximizing the covariance

$$\max_{\|w_s\|_2 = \|c_s\|_2 = 1} -t_s^T u_s = w_s^T X_s^T Y_s c_s$$

where $T$ means transpose and $\|\cdot\|_2$ denotes the $L_2$-norm. One solves the aforesaid problem by singular value decomposition (SVD) on $X_s^T Y_s$. Then, one computes the loading vector $p_s$ by the least square criterion as

$$\min_{p_s} \|X_s - t_s p_s^T\|_2^2$$

$$\Leftrightarrow p_s = X_s^T t_s / (t_s^T t_s)$$

thus organizing the regression from $t_s$ to $X_s$. PLSR supposes linear association from $t_s$ to $u_s$ furthermore by calculating a regression scalar $b_s$ by the least square criterion as

$$\min_{b_s} \|u_s - t_s b_s\|_2^2$$

$$\Leftrightarrow b_s = u_s^T t_s / (t_s^T t_s)$$

The residual matrices for the next iteration are updated

$$X_{s+1} = X_s - t_s p_s^T$$

$$Y_{s+1} = Y_s - b_s t_s c_s^T$$

Such procedures are repeated by PLSR for the optimal number of factors $S$, which is usually selected by cross-validation. One collects the outcomes from each iteration $T = [t_1, ..., t_S] \in \mathbb{R}^{L \times S}$, $P = [p_1, ..., p_S] \in \mathbb{R}^{N \times S}$, $B = \text{diag}(b_1, ..., b_S) \in \mathbb{R}^{S \times S}$, and $C = [c_1, ..., c_S] \in \mathbb{R}^{M \times S}$. As a result, one rewrites the decomposition of $X$ and the predicted response $\hat{Y}$ as

$$X = TP^T$$

$$\hat{Y} = TBC^T$$

Thus, the prediction relationship from $X$ to $\hat{Y}$ is structured as

$$\hat{Y} = XH$$

in which $H = P^T + BCT \in \mathbb{R}^{N \times M}$, and $P^T$ is the pseudo-inverse of $P^T$.

One notes that, the PLSR algorithm utilizes the least square criterion not only in (2)–(3) to build the regression relationship, but in (1) to calculate the projectors as well. Maximizing the covariance in (1) in essence could be rewritten as [30]

$$\min_{\|w_s\|_2 = \|c_s\|_2 = 1} \sum_{i=1}^{L} \left( \|x_s^l - x_s^l w_s w_s^T\|^2 + \|y_s^l - y_s^l c_s c_s^T\|^2 \right)$$

(7)

in which $x_s^l$ and $y_s^l$ are the $l$-th components for $X_s$ and $Y_s$, respectively. The meaning of (7) can be interpreted as follows. The first and second items in the summation are the quadratic reconstruction errors for the input and output, respectively. For example, $x_s^l w_s$ indicates the dimensionality-reduced input of the $l$-th observation, and $x_s^l w_s w_s^T$ denotes the reconstructed sample. The third item means the regression error between the latent variables. PLSR in each iteration finds the projectors $w_s$ and $c_s$ by minimizing the summation of three quadratic error items with the least square criterion, which is nonetheless non-robust with respect to deviated noises. Since eECoG recording is probably prone to physiological artifacts which exhibit large amplitudes [16], the decoding performance of PLSR could be deteriorated in practical application.

B. Regularized PLSR

Existing ameliorated versions of PLSR algorithm primarily exploit an additional regularization (penalization) item for the calculation of the prediction model. Specifically speaking, the regression relations that utilize the least square criterion can be reformulated by introducing regularization items. In [20], the $L_1$-norm regularization on the projectors was employed so as to compute sparse projectors, conducting the feature selection simultaneously. The $L_1$-regularized version rewrites (7) with $L_1$-regularization on the projectors by

$$\min_{\|w_s\|_2 = \|c_s\|_2 = 1} \sum_{i=1}^{L} \left( \|x_s^l - x_s^l w_s w_s^T\|^2 + \|y_s^l - y_s^l c_s c_s^T\|^2 \right)$$

$$\Leftrightarrow \sum_{i=1}^{L} \left( \|x_s^l - x_s^l w_s w_s^T\|^2 + \|y_s^l - y_s^l c_s c_s^T\|^2 \right)$$

$$\Leftrightarrow \lambda_1 \|w_s\|_1 + \lambda_2 \|c_s\|_1$$

(8)

in which $\lambda_1, \lambda_2 > 0$ denote the corresponding regularization parameters, and $\|\cdot\|_1$ means the $L_1$-norm operator. The authors further extended their study in [21], where Sobolev-norm and polynomial penalization were introduced into PLSR algorithm to strengthen the smoothness of the predicted output variable. Nevertheless, the regularized PLSR variants remain formulated based on the non-robust least-square foundation, thus probably being prone to performance degradation caused by the adverse noises.
III. PARTIAL MAXIMUM CORRENTROPY REGRESSION

A. Maximum Correntropy Criterion

The correntropy concept was developed in the field of ITL as a generalized correlation function of random processes [31], which measures the similarity and interaction between vectors in a kernel space. Correntropy associates with the information potential of quadratic Renyi’s entropy [23], in which the data’s probability density function (PDF) is estimated by the Parzen’s window method [32], [33]. The correntropy that evaluates the similarity between two arbitrary variables $A$ and $B$, which is denoted by $\mathcal{V}(A, B)$ in this paper, is

$$\mathcal{V}(A, B) = E[|k(A - B)|]$$

where $k(\cdot)$ is a kernel function satisfying the Mercer’s theory [34] and $E[\cdot]$ signifies the expectation operator. In the practical application, one calculates the correntropy with similarity between two arbitrary variables $A$ and $B$, which is denoted by $\mathcal{V}(A, B)$ in this paper, is

$$\hat{\mathcal{V}}(A, B) = \frac{1}{L} \sum_{l=1}^{L} g_{\sigma}(a_l - b_l)$$

where the Gaussian kernel function $g_{\sigma}(x) \triangleq \exp(-x^2/2\sigma^2)$ with kernel bandwidth $\sigma$ is the most widely used one for the kernel function $k(\cdot)$, thus leading to

$$\hat{\mathcal{V}}(A, B) = \frac{1}{L} \sum_{l=1}^{L} g_{\sigma}(a_l - b_l)$$

Maximizing the correntropy in (11), called as maximum correntropy criterion (MCC), exhibits numerous advantages. Correntropy is essentially a local similarity measure, that its correntropy criterion (MCC), exhibits numerous advantages. Moreover, it closely relates to the $m$-estimation, which can be regarded as a robust formulation of Welsch $m$-estimator [23], [35].

B. Partial Maximum Correntropy Regression

Substituting the three least-square items in the conventional PLSR [7] with the maximum correntropy yields

$$\max_{\|w_s\|_2 = 1, \|c_s\|_2 = 1} \sum_{l=1}^{L} \left( g_{\sigma_x}(x_l^T - x_l^T w_s w_s^T x_l^T) + g_{\sigma_y}(y_l^T - y_l^T c_s c_s^T y_l^T) + g_{\sigma_r}(x_l^T w_s - y_l^T c_s) \right)$$

in which $\sigma_x$, $\sigma_y$, and $\sigma_r$ are the Gaussian kernel bandwidths for $X_s$-reconstruction errors, $Y_s$-reconstruction errors, and the regression errors, respectively.

Then, one transforms the vectors $(x_l^T - x_l^T w_s w_s^T x_l^T)$ and $(y_l^T - y_l^T c_s c_s^T y_l^T)$ into scalars, provided that the two projects $w_s$ and $c_s$ are unit-length vectors, i.e.

$$\|w_s\|_2 = 1, \|c_s\|_2 = 1$$

Subsequently, one obtains the following optimization problem to acquire the projects $w_s$ and $c_s$ for the $s$-th iteration

$$\max_{\|w_s\|_2 = 1, \|c_s\|_2 = 1} \sum_{l=1}^{L} \left( g_{\sigma_x}(x_l^T x_l^T - x_l^T w_s w_s^T x_l^T) + g_{\sigma_y}(y_l^T y_l^T - y_l^T c_s c_s^T y_l^T) + g_{\sigma_r}(x_l^T w_s - y_l^T c_s) \right)$$

After obtaining $w_s$ and $c_s$, one calculates the latent variables as the conventional PLSR by $t_s = X_s w_s$ and $u_s = Y_s c_s$. We then compute the loading vector $p_s$ and the regression scalar $b_s$ by

$$\max_{p_s} \sum_{l=1}^{L} g_{\sigma_p}(x_l^T - t_l^T p_s)$$

$$\max_{b_s} \sum_{l=1}^{L} g_{\sigma_b}(u_l^T - t_l^T b_s)$$

in which $t_l^T$ and $u_l^T$ are the $l$-th elements for the latent variables $t_s$ and $u_s$, respectively. $\sigma_p$ and $\sigma_b$ are the corresponding kernel bandwidths. The residual matrices are updated by (4) as well.

One will repeat such procedures for the optimal number of factors and collects the acquired vectors from each iteration to organize the matrices $T$, $P$, $B$, and $C$, as the original PLSR. Ultimately, the predicted response $\hat{Y}$ can be obtained from $X$ by the regression relationship (6). The above-mentioned PLSR variant which is reformulated based on MCC, is called partial maximum correntropy regression (PMCR).

In what follows, we discuss about the optimization, convergence analysis, and determination of hyper-parameters considering the proposed PMCR algorithm.

1) Optimization: Three optimization problems (14) (15) (16) need to be addressed in PMCR. Those two regression problems (15) (16) could be well solved by the fixed-point optimization algorithm proposed in [36]. We mainly consider the problem (14) to calculate the robust projects $w_s$ and $c_s$. Based on the half-quadratic (HQ) optimization technique [27], [14] could be rewritten as

$$\max_{\|w_s\|_2 = 1, \|c_s\|_2 = 1} \sum_{l=1}^{L} \left( \sup_l \{\alpha_l(x_l^T x_l^T - x_l^T w_s w_s^T x_l^T) - \varphi(\alpha_l)\} + \sup_l \{\beta_l(y_l^T y_l^T - y_l^T c_s c_s^T y_l^T) - \varphi(\beta_l)\} \right.$$  

$$+ \sup_l \{\gamma_l(x_l^T w_s - y_l^T c_s))^2 \cdot \varphi(\gamma_l)\}$$

by introducing three sets of auxiliaries $\{\alpha_l\}_{l=1}^{L}$, $\{\beta_l\}_{l=1}^{L}$, and $\{\gamma_l\}_{l=1}^{L}$, respectively. Hence we conclude that optimizing (14) equals to optimizing $(\alpha_l, \beta_l, \gamma_l)$ and $(w_s, c_s)$ alternately by

$$\max_{\|w_s\|_2 = 1, \|c_s\|_2 = 1} \sum_{l=1}^{L} \left( \alpha_l x_l^T x_l^T - x_l^T w_s w_s^T x_l^T - \varphi(\alpha_l) \right.$$  

$$+ \beta_l y_l^T y_l^T - y_l^T c_s c_s^T y_l^T - \varphi(\beta_l) \right.$$  

$$+ \gamma_l(x_l^T w_s - y_l^T c_s))^2 \cdot \varphi(\gamma_l)$$

Since HQ method is also iterative, we denote the $s$-th factor in PMCR and $k$-th HQ optimization iteration by the subscript $(s, k)$. First, according to the HQ mechanism, we optimize the
Then, to optimize the projectors, we rewrite (18) by collecting the terms of projectors and omitting the auxiliaries as

\[
\begin{align*}
\alpha_{l,k+1} &= -\exp\left(-\frac{x_s^l x_s^{lT} - x_s^l w_{s,k}^l w_{s,k}^{lT} x_s^{lT}}{2\sigma_y^2}\right) \\
\beta_{l,k+1} &= -\exp\left(-\frac{y_s^l y_s^{lT} - y_s^l c_{s,k}^l c_{s,k}^{lT} y_s^{lT}}{2\sigma_y^2}\right) \\
\gamma_{l,k+1} &= -\exp\left(-\frac{(x_s^l w_{s,k} - y_s^l c_{s,k})^2}{2\sigma_y^2}\right)
\end{align*}
\]

(\(l = 1, \ldots, L\))

Then, to optimize the projectors, we rewrite (18) by collecting the terms of projectors and omitting the auxiliaries as

\[
\max \quad J_p = \sum_{l=1}^{L} \left( \left( \frac{\gamma_l}{\sigma_y^2} - \frac{\alpha_l}{\sigma_y^2}\right)x_s^l w_s^l x_s^{lT} + \left( \frac{\gamma_l}{\sigma_y^2} - \frac{\beta_l}{\sigma_y^2}\right)y_s^l c_s^l y_s^{lT} - \frac{\gamma_l}{\sigma_y^2} x_s^l w_s^l c_s^l y_s^{lT} \right)
\]

(20)

This exhibits a quadratic programming problem constrained by nonlinear equations, for which there exist numerous solutions in the literature, such as the sequential quadratic programming (SQP) that represents the state of the art in nonlinear programming methods [37]. The comprehensive procedures for PMCR are summarized in Algorithm 1.

Algorithm 1 Partial Maximum Correntropy Regression

1: \textbf{Input}: matrices of explanation X and response Y; number of factors S; a small positive value \(\varsigma\)
2: \textbf{Output}: regressor H for prediction \(\hat{Y} = XH\)
3: initialize \(X_1 = X\) and \(Y_1 = Y\);
4: for \(s = 1, 2, \ldots, S\) do
5: calculate the initial projectors \((w_s, c_s)\) by the conventional PLSR [1];
6: initialize \(converged = \text{FALSE}\);
7: repeat
8: \text{auxiliaries-step: update (\(\alpha_l, \beta_l, \gamma_l\)) with (19)};
9: \text{projectors-step: update \((w_s, c_s)\) with (20) by SQP;}
10: if the difference of the objective function (14) is smaller than \(\varsigma\) then
11: \(converged = \text{TRUE}\)
12: end if
13: until \(converged == \text{TRUE}\)
14: compute latent variables \(t_s = X_s w_s\) and \(u_s = Y_s c_s\);
15: compute \(p_s\) and \(b_s\) with MCC [15][16] with the fixed-point optimization algorithm proposed in [36];
16: update the residual matrices [4];
17: end for
18: organize the matrices \(T = \{t_1, \ldots, t_S\}\), \(P = \{p_1, \ldots, p_S\}\), \(B = \text{diag}(b_1, \ldots, b_S)\), and \(C = \{c_1, \ldots, c_S\}\);
19: compute the regressor \(H = P^T B C^T\);

2) Convergence Analysis: For the optimizations considering \(p_s\) and \(b_s\) which are addressed with the fixed-point algorithm proposed in [36], one finds the detailed convergence analysis in the cited literature. Here we mainly analyze the convergence of the projectors \(w_s\) and \(c_s\) in the optimization problem (14), which is addressed by the HQ technique. Since correntropy is in nature an \(m\)-estimator [23], the local optimums of (14) are supposed to be close to its global optimal solution, as proved in one theoretical study [38]. Accordingly, in what follows we analyze that (14) would converge to a local optimum with HQ optimization.

Proposition 1: If \(J_p(w_s,k, c_s,k) \leq J_p(w_{s,k+1}, c_{s,k+1})\) is established with fixing \((\alpha_l, \beta_l, \gamma_l) = (\alpha_{l,k+1}, \beta_{l,k+1}, \gamma_{l,k+1})\), the optimization problem with respect to the projectors in (14) would converge to a local optimum.

Proof: The convergence is proved as

\[
J(w_s,k, c_s,k, \alpha_l,k,l, \beta_l,k,l, \gamma_l,k,l) \leq J(w_{s,k+1}, c_{s,k+1}, \alpha_{l,k+1}, \beta_{l,k+1}, \gamma_{l,k+1})
\]

(21)

where the first inequality is guaranteed by the HQ mechanism [27], while the second inequality arises from the assumption of the present proposition.

One realizes consequently that, to achieve the convergence of the projectors in (14), it is unnecessary to strictly attain the maximum of (20) at every processors-step of Algorithm 1. On the contrary, so long as the updated projectors exhibit a larger objective function \(J_p\) at each processors-step, the problem (14) will converge as proved above. This leads to great convenience in practice, that one only needs a few SQP iterations. One can finish the processors-step once confirming the increase on \(J_p\), thus accelerating the convergence in practical implementation.

3) Hyper-Parameter Determination: Notably, there exist five Gaussian kernel bandwidths \(\sigma_x\), \(\sigma_y\), \(\sigma_r\), \(\sigma_p\), and \(\sigma_b\), respectively, to be determined for practicable implementation. In the literature, an effective approach to acquire a rather appropriate kernel bandwidth for probability density estimation, named as Silverman’s rule, was proposed in [32]. Denoting the current set of errors by \(E\) with \(L\) observations, the bandwidth is

\[
\sigma^2 = 1.06 \times \min\{\sigma_E, \frac{R}{1.34}\} \times (L)^{-1/5}
\]

(22)

in which \(\sigma_E\) is the standard deviation of the \(L\) errors, and \(R\) denotes the interquartile range.

IV. EXPERIMENTS

In the present part, we evaluate the proposed PMCR method on synthetic dataset and real eECoG dataset, respectively, with comparing it to existing PLSR methods. Specifically speaking, we compare the proposed PMCR with the conventional PLSR method and the state-of-the-art L1-regularized PLSR [20]. For evenhanded comparison, all the evaluated methods employ an identical number of factors, which is selected by the original PLSR algorithm in five-fold cross-validation, with the maximal number of factors being 100.

Considering the performance indicators for algorithm evaluation, we use three typical measures in regression: i) Pearson’s correlation coefficient (\(r\)), ii) root mean squared error (RMSE) which is computed by

\[
\text{RMSE} = \sqrt{\frac{1}{L} \sum_{l=1}^{L} ||\hat{y}_l - y_l||^2}
\]

(23)
in which $\hat{y}_l$ and $y_l$ are the $l$-th observations for the prediction $\hat{Y}$ and the target $Y$, respectively, and iii) mean absolute error (MAE) which represents the average $L_1$-norm distance

$$\text{MAE} = \frac{1}{L} \sum_{i=1}^{L} \| \hat{y}_i - y_i \|$$ (24)

One notes that, to compare the robustness between different algorithms, only contaminating the training samples by noises with isolating testing data from contamination is an extensively approved and implemented method in the literature, as advised in [39]. Accordingly, only the training samples suffer adverse contamination in the following experiments.

A. Synthetic Dataset

First, we consider an inter-correlated, high-dimensional, and noisy synthetic dataset, with which various PLSR methods are assessed under different levels of contamination. Randomly we create 300 i.i.d. latent variables $t \sim U(0, 1)$ for training, and 300 i.i.d. latent variables $t' \sim U(0, 1)$ for testing, where $U$ is uniform distribution and the dimension of $t$ is 20. We generate the hypotheses from the latent variable to the explanatory and response then. Specifically, we form the transformations with arbitrary values subject to the standard normal distribution. $t$ is multiplied with an $\mathbb{R}^{20 \times 500}$ matrix to acquire $x$, and a 20-dimensional vector to obtain $y$. As a result, an inter-correlated and high-dimensional dataset is produced, in which we predict the univariate response from 500-dimensional observations by 300 training samples, and evaluate the prediction performance on the other 300 testing samples.

Considering the contamination on the synthetic dataset, we suppose the explanatory matrix $X$ to be contaminated, because the adverse noise mainly happens to the brain recording which is usually employed as the explanatory in BCI systems. To this end, a part of training samples are stochastically selected from the training set, the inputs of which are then replaced by noises with large amplitudes. The noise level signifies the proportion of the contaminated samples in the entirety, which is increased from 0 to 1.0 with step 0.05. For the distribution of noise, we utilize the zero-mean Gaussian distribution with large standard deviations so as to imitate outliers, in which 30, 100, and 300 are employed, respectively. We assess the different algorithms with 100 Monte-Carlo repetitive trials, and present the average results under the three performance indicators in Fig. I.

One observes from Fig. I that for all the three different noise standard deviations 30, 100, and 300, the proposed PMCR can realize superior regression performance compared to the other existing PLSR algorithms consistently for $r$, RMSE, and MAE, respectively, when the synthetic dataset is contaminated.

B. eECoG Dataset

To further demonstrate the superior robustness of the PMCR algorithm, we evaluate the different PLSR algorithms with the following practical brain decoding task. In this subsection, we employ the publicly available Neurotycho eECoG dataset that was initially proposed in [16]. Two adult Japanese macaques, denoted by $B$ and $C$, respectively, were commanded to capture the foods with their right hands, in which the continuous three-dimensional hand trajectories of 120 Hz sampling rate and 64-channel eECoG of 1,000 Hz sampling rate on the contralateral hemisphere were recorded simultaneously. For both Macaque $B$ and $C$, ten recordings of 15-minute duration were executed, in which one trains a prediction model on the data of the first 10 minutes in each recording, and then evaluates the decoding performance on the remaining 5 minutes of the corresponding recording.

Considering feature extraction, we employ the identical offline decoding paradigm as in [16]. Initially, eECoG signals are pre-processed by a ten-order Butterworth band-pass filter with cutoff frequencies from 1 to 400 Hz, and then re-referenced by common average referencing (CAR) approach. The continuous three-dimensional trajectories of right wrist are down-sampled to 3 Hz, thus, resulting in 2,700 observations in each recording (3 Hz × 60 sec × 15 min). One then predicts the three-dimensional position of time $t$ from the eECog signals in the previous one second. Time–frequency representation is exploited to describe the eECoG feature. To predict the position of time $t$, eECoG of each channel from $t-1.1s$ to $t$ is processed with Morlet wavelet transformation at ten center frequencies from 10 to 120 Hz. The time–frequency spectrum is then resampled at 10 temporal lags ($t-1s$, $t-0.9s$, $t-0.8s$, $t-0.7s$). Accordingly, each observation exhibits a 6,400-dimensional feature (64 channels × 10 frequencies × 10 temporal lags). As a result, one trains a regression model with 1,800 samples of 6,400-dimensional features, and predicts the remaining 900 trajectories.

To assess the robustness of various methods, eECoG signals are contaminated artificially with deviated outliers to simulate detrimental artifacts. Specifically speaking, we select a certain proportion of the training eECoG samplings stochastically and deteriorate them with noises which obey a zero-mean Gaussian distribution with a variance 20 times the one of corresponding channel. The illustrative paradigms for eECoG signal decoding are summarized in Fig. II. Note that, the ‘Noise Level’ signifies the ratio of the contaminated eECoG samplings in this section, as opposed to that of the deteriorated ones out the entire 1,800 training samples. The ratio of the affected training samples can be evidently larger than the indicated noise level, because one contaminated ECoG sampling will infect several time windows for the time–frequency feature calculation. For example, when the noise level is indicated as 10$^{-3}$, the deteriorated proportion of the training set is $(0.6645±0.0089)$. Moreover, we illustrate how the impulsive eECoG sampling noises influence the time–frequency features in Fig. III where one perceives heavy-tailed characteristic on the feature noises, which is intractable for the least square criterion. In addition, the effects of high-frequency range are more notable, due to the property of impulsive noise.

For both Macaque $B$ and $C$, we implement 10 Monte-Carlo repetitive trials for each recording under different noise levels, thus evaluating each method 100 times for both macaques with each noise level. The resultant performance indicators are then summarized in TABLE VI where the optimal results under each condition are marked in bold. One finds from TABLE VI that the proposed PMCR algorithm realizes the optimal regression performance under most conditions, outperforming the other existing PLSR algorithms. To be specific, the proposed PMCR
**Fig. 1**: Average regression performance indicators of the inter-correlated, high-dimensional, and contaminated synthetic dataset under different noise standard deviations with noise levels from 0 to 1.0: (a): noise standard deviation = 30, (b): noise standard deviation = 100, and (c): noise standard deviation = 300.

**Fig. 2**: Experimental protocol of Neurotycho epidural ECoG dataset and decoding schema to evaluate the robustness of different PLSR algorithms. (a) Macaques retrieved foods in a three-dimensional random location, in which the body-centered coordinates of right wrist and epidural ECoG signals were recorded simultaneously. (b) Macaque B and C were implanted with 64-channel epidural ECoG electrodes on the contralateral (left) hemisphere, overlaying the regions from prefrontal cortex to parietal cortex. Ps: principal sulcus, As: arcuate sulcus, Cs: central sulcus, and IPs: intraparietal sulcus. Notably, (a)(b) are excerpted from [16] which provides the public dataset descriptions in http://neurotycho.org/epidural-ecog-food-tracking-task. (c) Decoding diagram from eECoG signal to continuous hand trajectory. The training eECoG signals are contaminated artificially with different noise levels to evaluate the robustness of various PLSR algorithms.

achieves the optimal consequences in 5, 6, and 7 out of 8 noise levels for X, Y, and Z coordinates of Macaque B, respectively, and 6, 6, and 6 out of 8 noise levels for X, Y, and Z coordinates of Macaque C, respectively, exhibiting promising prediction competence in the practical decoding tasks from noisy eECoG signals to three-dimensional continuous hand trajectories.

Studying how spatio-spectro-temporal weights contribute to the prediction models helps investigate the neurophysiological patterns. The component of the learned prediction models $H$ by different PLSR algorithms is denoted by $h_{ch, freq, temp}$, which is corresponding to the electrode ‘ch’, the frequency ‘freq’, and the temporal lag ‘temp’. Thus one could calculate the spatio-spectro-temporal contributions by quantifying the ratio of the summation of absolute value considering the present domain.
to the absolute value summation of the entire prediction model

\[ W_c(ch) = \frac{\sum_{freq} \sum_{temp} |h_{ch,freq,temp}|}{\sum_{ch} \sum_{freq} \sum_{temp} |h_{ch,freq,temp}|} \]  

\[ W_f(freq) = \frac{\sum_{ch} \sum_{freq} \sum_{temp} |h_{ch,freq,temp}|}{\sum_{ch} \sum_{freq} \sum_{temp} |h_{ch,freq,temp}|} \]  

\[ W_t(temp) = \frac{\sum_{ch} \sum_{freq} \sum_{temp} |h_{ch,freq,temp}|}{\sum_{ch} \sum_{freq} \sum_{temp} |h_{ch,freq,temp}|} \]

(25)  

(26)  

(27)  

To demonstrate the robustness of PMCR to extract more accurate neurophysiological information, we compare the spatio-spectro-temporal patterns acquired by different methods from uncontaminated and contaminated (noise level = 10^{-2}) eECoG signals in Fig. 4.

One can observe from Fig. 4 that, when the eECoG signal is contaminated, the proposed PMCR acquires neurophysiological patterns with better consistency. For Macaque B, compared to the spatial patterns acquired from the acoustic signal which exhibit evident concentration in dorsal premotor cortex (PMD), the patterns acquired by different methods from contaminated eECoG all show varying degrees of concentration, in which PMCR reveals the maximal contribution in PMD among the affected spatial patterns. For spectral contributions, Macaque B exhibits two contribution peaks at 40Hz and 69Hz with acoustic signal. PMCR obtains the highest contributions at these two frequency bins in the three evaluated methods. Considering the temporal contribution, which discloses the largest concentration at -0.1s without contamination for Macaque B, PMCR achieves higher contribution at -0.1s than other algorithms. One could observe similar occurrence from Macaque C.

V. DISCUSSION

A. NUMBER OF FACTORS

The number of factors \( S \) plays a vital role in PLSR methods, representing the iteration numbers to decompose the input and output matrices. Since it usually causes a notable effect on the result, additionally, we assess the performance with respect to the numbers of factors for each method. To this end, we exploit the synthetic dataset from Subsection IV-A with noise standard deviation being 100 under three different noise levels, 0.3, 0.6, and 0.9, respectively. The resultant performance indicators of each algorithm with respect to the numbers of factors from 1 to 100 are illustrated in Fig. 5 with 100 repetitive trials.

One perceives from Fig. 5 that not only the proposed PMCR eventually achieves superior regression performance indicators with the optimal number of factors, but it could realize a rather acceptable performance with a small number of factors as well. Specifically speaking, when the noise level is equal to 0.3 and 0.6, PMCR achieves high correlation coefficients consistently, and exhibits significant descent on RMSE and MAE when the number of factors is smaller than 10. When the noise level is 0.9, PMCR shows consistency on three performance indicators with the number of factors being larger than 20, whereas the other methods acquire their optimal performance with a larger number of factors. This suggests that, the proposed PMCR can abstract sufficient information with a small number of factors from training samples in a noisy regression task, demonstrating the effectiveness of reformulating PLSR with MCC.

B. PMCR WITH REGULARIZATION

One should additionally note that, the PMCR was proposed by reformulating the conventional PLSR algorithm with using the robust MCC, instead of the mediocre least square criterion. Thus, the proposed PMCR exhibits the supplementary potential for further performance improvements with regularization terms, the same as in the existing regularized PLSR methods. For instance, the \( L_1 \)-norm regularization can be employed in (4) to encourage sparse and robust projectors. In addition, if one requires better smoothness on the predictions, polynomial or Sobolev-norm penalization could be introduced into PMCR. MCC-based algorithm with regularization has been examined in the literature. For example, a robust implementation of sparse representation classifier (SRC) for face recognition was proposed by merging MCC and \( L_1 \)-norm regularization [40].

C. EXTENSION TO MULTI-WAY APPLICATION

Similarly to the generic PLSR algorithm which projects the data into the low-dimensional space of latent variables, multi-way PLSR structures the relationship for tensor variables with dimensionality reduction technique by tensor factorization. As reported in the literature, multi-way PLSR can usually exhibit better prediction performance than the generic PLSR algorithm.
in brain decoding tasks, where one is supposed to organize the spatio-spectro-temporal feature with a tensor form. Essentially, multi-way PLSR decomposes the input and output tensors with the least square criterion by minimizing the Frobenius norm, which represents a generalization of $L_2$-norm [41].

The proposed PMCR method treats the regression problem of matrix variable, i.e. two-way variable. Extending the PMCR algorithm to multi-way application could probably improve the prediction performance further, which would be investigated in depth in our future works. Promisingly, MCC has been verified effective for tensor variable analysis in a recent study [42].

### VI. CONCLUSION

This paper proposes a new robust variant of PLSR algorithm by reformulating the non-robust least square criterion with the sophisticated MCC framework. The proposed robust objective functions can be efficaciously optimized by half-quadratic and fixed-point-based optimization means. Extensive experimental results on the synthetic dataset and Neurotycho epidural EEG dataset, respectively, demonstrate that the proposed PMCR can outperform the conventional and regularized PLSR algorithms, exhibiting promising robustness in high-dimensional and noisy brain decoding tasks.
Fig. 4: Spatio-spectro-temporal contributions of different decoding algorithms averaged across X-, Y-, and Z-coordinates under noise levels 0 and 0.01. (a) Spatial contributions for each electrode. (b) Spectral contributions for each frequency. (c) Temporal contributions for each temporal lag.

REFERENCES

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, “Brain–computer interfaces for communication and control,” Clinical neurophysiology, vol. 113, no. 6, pp. 767–791, 2002.
[2] J. P. Donoghue, “Connecting cortex to machines: recent advances in brain interfaces,” Nature neuroscience, vol. 5, no. 11, pp. 1085–1088, 2002.
[3] F. A. Mussa-Ivaldi and L. E. Miller, “Brain–machine interfaces: computational demands and clinical needs meet basic neuroscience,” TRENDS in Neurosciences, vol. 26, no. 6, pp. 329–334, 2003.
[4] M. A. Lebedev and M. A. Nicolelis, “Brain–machine interfaces: past, present and future,” TRENDS in Neurosciences, vol. 29, no. 9, pp. 536–546, 2006.
[5] S. Amiri, R. Fazel-Rezai, and V. Asadpour, “A review of hybrid brain-computer interface systems,” Advances in Human-Computer Interaction, vol. 2013, 2013.
[6] C. A. Chestek, A. P. Batista, G. Santhanam, M. Y. Byron, A. Afshar, J. P. Cunningham, V. Gilja, S. I. Ryu, M. M. Churchland, and K. V. Shenoy, “Single-neuron stability during repeated reaching in macaque premotor cortex,” Journal of Neuroscience, vol. 27, no. 40, pp. 10742–10750, 2007.
[7] G. Buzsáki, C. A. Anastassiou, and C. Koch, “The origin of extracellular fields and currents—eeg, ecog, lfp and spikes,” Nature reviews neuroscience, vol. 13, no. 6, pp. 407–420, 2012.
[8] S. P. Levine, J. E. Huggins, S. L. BeMent, R. K. Kushwaha, L. A. Schuh, M. M. Rohde, E. A. Passaro, D. A. Ross, K. V. Eliseevich, and B. J. Smith, “A direct brain interface based on event-related potentials,” IEEE Transactions on Rehabilitation Engineering, vol. 8, no. 2, pp. 180–185, 2000.
[9] E. C. Leuthardt, G. Schalk, J. R. Wolpaw, J. G. Ojemann, and D. W. Moran, “A brain–computer interface using electrocorticographic signals in humans,” Journal of neural engineering, vol. 1, no. 2, p. 63, 2004.
[10] T. Ball, A. Schulze-Bonhage, A. Aertsen, and C. Mehring, “Differential representation of arm movement direction in relation to cortical anatomy and function,” Journal of neural engineering, vol. 6, no. 1, p. 016006, 2009.
[11] E. C. Leuthardt, K. J. Miller, G. Schalk, R. P. Rao, and J. G. Ojemann, “Electrocorticography-based brain computer interface-the seattle experience,” IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 14, no. 2, pp. 194–198, 2006.
Fig. 5: Average regression performance indicators of the synthetic dataset with noise standard deviation 100 under three noise levels, 0.3, 0.6, and 0.9, respectively, and different numbers of factors from 1 to 100.