Multi-way Feature Selection for ECoG-based Hand Movement Prediction

Anastasia Motrenko[*], Vadim Strijov[†]

The paper addresses the problem of feature selection in regression models in application to ECoG-based motion decoding. The task is to predict hand trajectories from the voltage time series of cortical activity. Feature description of a each point resides in spatial-temporal-frequency domain and include the voltage time series themselves and their spectral characteristics. Feature selection is crucial for adequate solution of this regression problem, since electrocorticographic data is highly dimensional and the measurements are correlated both in time and space domains. We propose a multi-way formulation of QPFS, a recent approach to filtering-based feature selection [1]. QPFS incorporates both estimates of similarity between features, and their relevance to the regression problem, and allows an effective way to leverage them by solving a quadratic program. Our modification allows to apply this approach to multi-way data. We show that this modification improves prediction quality of resultant models.

Keywords: *feature selection, brain-computer interface, decoding electrocorticographic data, multi-way data, partial least squares regression*

1 Introduction

Analysis of cortical activity during motor imagery is essential for designing Brain Computer Interfaces. The goal of motor imagery analysis is to recognize intended movements from the recorded brain activity. While there are various techniques for measuring cortical data for BCI [2, 3], we concentrate on the ElectroCorticoGraphic (ECoG) signals [4]. ECoG, as well as other invasive techniques provide more stable recordings and better better resolution in temporal and spatial domains, compared to their non-invasive counterparts.

We address the problem of continuous trajectory reconstruction. The subdural ECoG signals are measured across 32 channels as the subject is moving a limb [5] and then transformed into feature vectors for selected time points. Given a transformation of the ECoG time series into informative features, the problem of trajectory reconstruction is a regression problem. A common approach to feature extraction is to apply continuous wavelets transform features to the ECoG time series. Since the resulting spatial-temporal-spectral representation is highly redundant, various feature selection and dimensionality reduction techniques are used [6, 7] to extract only the most relevant features.

A widely used technique is PLS [6, 8, 9] and its extensions for multi-way data [4, 6]. Multi-way representation is actively used in analysis of biomaterial and chemical data due to the multi-way structure of the data in this domains. The idea is that unfolding data to matrices might lead to neglecting important dependencies present in the unfolded dimension of the multi-way data.

[*] anastasia.motrenko@phystech.edu, Moscow Institute of Physics and Technology

[†] Dorodnicyn Computing Center of RAS

Contrarily, using multi-way approaches preserves the data structure, which may help to improve regression quality, as demonstrated in [6]. Similarly to the original PLS and SVD, multi-way extensions of PLS rely on multi-way decompositions usually Tucker or PARAFAC [10]. Since PLS is prone to overfitting, several regularisation techniques were proposed to increase its stability [4].

For comfortable BCI usage, the regression algorithms must have minimum latency. Computational efficiency is one of the reasons why PLS is so frequently used in ECoG-based scalogram to coordinates regression. As a feature selection method, PLS classifies as the embedded method, which means it performs dimensionality reduction and parameter estimation simultaneously. An alternative approach is to used filter feature selection methods, which are model-free and allow to select features without the need to actually train the model. A recent approach to filtering feature selection by Katrutsa [1] is formulated as a quadratic program which minimizes correlation between features while maximizing feature relevance. The drawback here is that model is not taken into account. Since the original QPFS does not consider multiway structure of the data, we formulate a multiway extension of QPFS. The compare the original and multiway QPFS applied to trajectory reconstruction problem and show that proposed modification improves prediction quality.

2 Problem statement

The raw ECoG data contains multivariate time series $\mathbf{s}(t) \in \mathbb{R}^{N_{\text{ch}}}$ with voltage measurements for each channel $1, \dots, N_{\text{ch}}$, and multivariate target time series $\mathbf{y}(t) \in \mathbb{R}^3$ with 3D limb coordinates. These time series are converted to the data sample $(\underline{\mathbf{D}}, \mathbf{Y})$:

$$\underline{\mathbf{D}} \in \mathbb{R}^{T \times F \times N_{\text{ch}} \times M}, \; D_{(m,:,:,:)} = \underline{\mathbf{X}}_m, \quad \mathbf{Y} = [\mathbf{y}(t_1)^\intercal, \dots, \mathbf{y}(t_M)^\intercal]^\intercal,$$

such that $\mathbf{y}_m = \mathbf{y}(t_m)$ and $\underline{\mathbf{X}}_m \in \mathbb{R}^{T \times F \times N_{\text{ch}}}$ is a three-way matrix. Each slice $\underline{\mathbf{X}}_m^{(:,:,n)} \in \mathbb{R}^{T \times F}$ of $\underline{\mathbf{X}}_m$ stores time-frequency features extracted from the time series $[s_n(t_m - \Delta t), \dots, s_n(t))]$ along the channel n, $n = 1, \dots, N_{\text{ch}}$. The procedure of feature extraction $\mathbf{s}(t) \to \underline{\mathbf{X}}_m$ will be described in more detail in the Appendix.

The problem is to reconstruct the trajectory $\mathbf{Y}$ of each limb given $\underline{\mathbf{X}}_m$, $m = 1, \dots M$. The reconstructed trajectory $\hat{\mathbf{Y}}$ approximates the real one as a linear combination of features:

$$\hat{\mathbf{y}}_m = \text{vec}\big(\underline{\mathbf{X}}_m\big)^\intercal \hat{\mathbf{w}}, \tag{1}$$

where the weight vector $\hat{\mathbf{w}} \in \mathbb{R}^{T \cdot F \cdot N_{\text{ch}} \times 3}$ minimize the squared sum of residues:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} ||\hat{\mathbf{Y}} - \mathbf{Y}||_2^2. \tag{2}$$

Feature selection. Due to the fact that ECoG measurements are highly correlated both in time and space, the problem (1), (2) is instable. To decrease computational cost and increase stability of reconstruction we use feature selection methods and add regularization term to the problem (2).

Let $\mathbf{X} \in \mathbb{R}^{M \times T \cdot F \cdot N_{\text{ch}}}$ denote vectorized feature matrix $\underline{\mathbf{D}} \in \mathbb{R}^{T \times F \times N_{\text{ch}} \times M}$:

$$\mathbf{X} = [\text{vec}(\underline{\mathbf{X}}_1), \dots, \text{vec}(\underline{\mathbf{X}}_M)]^\intercal =$$

$$[\dots, \boldsymbol{\chi}_{ijk}, \dots], \; (i, j, k) \in \{1, \dots, T\} \times \{1, \dots, F\} \times \{1, \dots, N_{\text{ch}}\}.$$

Define an indicator variable $\underline{\mathbf{A}} \in \mathbb{R}^{T\times F\times N_{\text{ch}}}$, which encodes inclusions of features $\boldsymbol{\chi}_{ijk}$ into the dataset and the corresponding two-way feature matrix:

$$\mathbf{X}_{\underline{\mathbf{A}}} = [\dots, \boldsymbol{\chi}_{ijk}, \dots], \text{ such that } \underline{\mathbf{A}}_{ijn} = 1.$$

Feature selection problem is formulated the following way:

$$\underline{\mathbf{A}} = \underset{\underline{\mathbf{A}}\in\mathbb{R}^{T\times F\times N_{\text{ch}}}}{\arg\min}\ \mathcal{L}\left(\mathbf{X}_{\underline{\mathbf{A}}}\mathbf{w}_{\underline{\mathbf{A}}}, \mathbf{Y}\right),$$

where $\mathcal{L}(\hat{Y}, \mathbf{Y})$ is some loss function and $\mathbf{w}_{\underline{\mathbf{A}}}$ minimizes quadratic loss (2) for $\mathbf{X}_{\underline{\mathbf{A}}}$.

To evaluate forecasting quality, we use scaled MSE

$$\text{sMSE}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{\sum_{m=1}^{M} ||\hat{\mathbf{y}}_m - \mathbf{y}_m||_2}{\sum_{m=1}^{M} ||\bar{\mathbf{y}} - \mathbf{y}_m||_2}, \quad \bar{\mathbf{y}} = \frac{1}{M}\sum_{i=1}^{M} y_m, \tag{3}$$

and correlation coefficient between predictions $\hat{Y}$ and the original data $\mathbf{Y}$:

$$\text{corr}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{\text{cov}(\hat{\mathbf{y}}, \mathbf{y})}{\sqrt{\text{cov}(\hat{\mathbf{y}}, \hat{\mathbf{y}})\text{cov}(\mathbf{y}, \mathbf{y})}}, \tag{4}$$

2.1 Quadratic Programming Feature Selection.

The features are correlated in time, space and frequency domains. To reduce redundancy of feature description and increase stability of model we apply feature selection. We consider a filtering feature selection approach, proposed in [1]. Filtering approaches, which assign scores to each variable, are generally more fast than embedded or wrapper approaches. However, since they do not consider relationships between variables, filtering methods tend to select correlated features. The advantage of quadratic programming feature selection (QPFS) technique, proposed in [1] is that it considers both relevance and similarity between features without looking at all subsets of features. The feature selection problem is formulated as quadratic programming problem

$$\mathbf{a} = \underset{\mathbf{a}\in\{0,1\}^N}{\arg\min}\left(\mathbf{a}^\mathsf{T}\mathbf{Q}\mathbf{a} - \mathbf{b}^\mathsf{T}\mathbf{a}\right), \tag{5}$$

where q_{ij} entry of matrix $\mathbf{Q} \in \mathbb{R}^{N\times N}$ quantifies *similarity* between i-th and j-th features, say

$$q_{ij} = |\text{corr}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)|.$$

Here $\boldsymbol{\chi}_i$, $\boldsymbol{\chi}_j$ denote columns of the design matrix $\mathbf{X}$. Similarly, element b_i, which is referred to as *relevance* of the $i-$th feature, quantifies similarity between $\boldsymbol{\chi}_i$ and the target $\mathbf{Y}$:

$$b_i = \frac{1}{3}\sum_{n=1}^{3} |\text{corr}(\boldsymbol{\chi}_i, \mathbf{y}_n)|. \tag{6}$$

Note that there are other ways to define $\mathbf{Q}$ and $\mathbf{b}$ besides the correlation coefficient. For example, [1] also considers mutual information and normalized feature significance as similarity and relevance measures.

The problem (5) balances similarity between selected features and their predictive importance through optimization by a binary vector $\mathbf{a} \in \mathbb{R}^N$, which defines the active set of predictors:

$$\mathbf{X} = [\boldsymbol{\chi}_{i_1}, \dots, \boldsymbol{\chi}_{i_n}], \text{ where } a_{i_k} = 1,\ k = 1, \dots, n.$$

Multi-way QPFS. The problem (5) is formulated for the two-way data. In case of four-way[1] data $\underline{\mathbf{D}}$ the predictors $\underline{\mathbf{X}}_m$ and the indicator variable $\underline{\mathbf{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ are three-way matrices. Instead of feature vector $\boldsymbol{\chi}_i \in \mathbb{R}^n$ we now have $\boldsymbol{\chi}_{ijk} = \underline{\mathbf{D}}_{(:,i,j,k)} \in \mathbb{R}^{n_1 n_2 n_3}$. Finally, the relevance matrix is the same size as $\underline{\mathbf{X}}_m$ and similarities between features are represented by two-way matrices $\mathbf{Q}_1 \in \mathbb{R}^{n_1 \times n_1}$, $\mathbf{Q}_2 \in \mathbb{R}^{n_2 \times n_2}$, $\mathbf{Q}_3 \in \mathbb{R}^{n_3 \times n_3}$, one for each mode.

To incorporate the multi-way structure of ECoG features, we propose a multi-way formulation of QPFS. Let $\mathbf{a} \circ \mathbf{b}$ denote the outer product of two vectors $\mathbf{a} \in \mathbb{R}^{n_1}, \mathbf{b} \in \mathbb{R}^{n_2}$

$$\mathbf{a} \circ \mathbf{b} \in \mathbb{R}^{n_1 \times n_2} : [\mathbf{a} \circ \mathbf{b}]_{ij} = a_i b_j, \quad \mathbf{a} \in \mathbb{R}^{n_1}, \mathbf{b} \in \mathbb{R}^{n_2},$$

$\underline{\mathbf{A}} \times_d \mathbf{B}$ denote the d-mode product of multi-way matrix $\underline{\mathbf{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ to matrix $\mathbf{B} \in \mathbb{R}^{m \times n_1}$

$$\underline{\mathbf{A}} \times_1 \mathbf{B} \in \mathbb{R}^{m \times n_2 \times n_3} : [\underline{\mathbf{A}} \times_1 \mathbf{B}]_{ijk} = \sum_{i'} a_{i'jk} b_{ii'},$$

and $\underline{\mathbf{A}} * \underline{\mathbf{B}}$ denote the element-wise product:

$$[\underline{\mathbf{A}} * \underline{\mathbf{B}}]_{ijk} = a_{ijk} b_{ikj}.$$

]

Suppose the similarity matrices $\mathbf{Q}_1 \in \mathbb{R}^{n_1 \times n_1}$, $\mathbf{Q}_2 \in \mathbb{R}^{n_2 \times n_2}$, $\mathbf{Q}_3 \in \mathbb{R}^{n_3 \times n_3}$ for each mode of $\underline{\mathbf{X}}$ and a multi-way relevance $\underline{\mathbf{B}}$ matrix are known. The problem (5) reformulates as follows:

$$\underline{\mathbf{A}} = \operatorname*{arg\,min}_{\underline{\mathbf{A}} \in \{0,1\}^{n_1 \times n_2 \times n_3}} \left(\sum_{d=1}^{3} (\underline{\mathbf{A}} \times_1 \mathbf{Q}_d) * \underline{\mathbf{A}} + \underline{\mathbf{B}} * \underline{\mathbf{A}} \right) \times_1 \mathbf{1}_{n_1} \times_2 \mathbf{1}_{n_2} \times_3 \mathbf{1}_{n_3}, \tag{7}$$

Note that operation $\underline{\mathbf{A}} \times_1 \mathbf{1}_{n_1} \times_2 \mathbf{1}_{n_2} \times_3 \mathbf{1}_{n_3}$ is equivalent to summation over all entries of $\underline{\mathbf{A}}$:

$$\underline{\mathbf{A}} \times_1 \mathbf{1}_{n_1} \times_2 \mathbf{1}_{n_2} \times_3 \mathbf{1}_{n_3} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_3} \sum_{k=1}^{n_3} a_{ijk}.$$

Solution of (7) is based on low rank decomposition of $\underline{\mathbf{A}}$, which allows to solve the problem (7) solves iteratively, via alternate approach, so that at each step a quadratic programming problem is solved. The derivation and exact formulation of the multi-way QPFS algorithms can be found in Section 6.

Similarity and relevance for multi-way data. To define d-mode similarity matrix $\mathbf{Q}_d$, $d = 1, 2, 3$ we use higher-order SVD decomposition (PARAFAC):

$$\underline{\mathbf{X}} = \sum_{r=1}^{R} \lambda_r \mathbf{u}_0^{(r)} \circ \mathbf{u}_1^{(r)} \circ \mathbf{u}_2^{(r)} \circ \mathbf{u}_3^{(r)}.$$

The d-mode similarities $\mathbf{Q}_d$ are computed as:

$$\mathbf{Q}_d = \frac{1}{R-1} \mathbf{U}_d \Sigma \mathbf{U}_d^{\mathsf{T}}, \quad \text{where } \Sigma = \operatorname{diag}(\lambda_1, \dots, \lambda_R),$$

$$\mathbf{U}_d = [\mathbf{u}_d^{(1)}, \dots, \mathbf{u}_d^{(R)}] \in \mathbb{R}^{n_d \times R}, \ d = 1, 2, 3.$$

The relevance definition (6) generalizes straightforwardly to the three-way case:

$$\underline{\mathbf{B}} = [b_{ijk}], \quad b_{ijk} = \frac{1}{3} \sum_{n=1}^{3} |\operatorname{corr}(\boldsymbol{\chi}_{ijk}, \mathbf{y}_n)|.$$

[1] We formulate the multi-way QPFS for the case of four-way data (three-way features), but all the derivations generalize to other number of modes $\underline{\mathbf{X}}_m \in \mathbb{R}^{n_1 \times \cdots \times n_d}, d \geq 2$.

Linear relaxation of (7)**.** The problem (7) is the integer optimization problem, which is not convex. To allow for more efficient solution, we have to relax $\underline{\mathbf{A}} \in \{0,1\}^{n_1\times n_2\times n_3}$ constraint into $\underline{\mathbf{A}} \in [0,1]^{n_1\times n_2\times n_3}$. After the solution $\hat{\underline{\mathbf{A}}}$ of the relaxed problem is found, we threshold $\hat{\underline{\mathbf{A}}}$

$$\underline{\mathbf{A}}(\epsilon) = [a_{ijk}], \quad a_{ijk} = \begin{cases} 1 \text{ if } \hat{a}_{ijk} \geq \epsilon, \\ 0 \text{ otherwise.} \end{cases} \tag{8}$$

to select a number of features $\mathbf{X}_{\underline{\mathbf{A}}}$. Setting various threshold values ϵ, we obtain various active sets of features $\mathbf{X}_{\underline{\mathbf{A}}(\epsilon)}$. Solution $\hat{\underline{\mathbf{A}}}$ of relaxed QPFS defines order on the feature set: $\boldsymbol{\chi}_{ijk} \preceq \boldsymbol{\chi}_{i'j'k'} \Leftrightarrow \hat{a}_{ijk} \leq \hat{a}_{i'j'k'}$.

2.2 Partial Least Squares regression

We use PLS as the alternative to the proposed feature selection. Instead of selecting features from the original feature space, PLS reduces its dimensionality by selecting several factors — linear combinations of the original features. An attractive feature of PLS is that it does so with regard to the targets, so that the resultant factors are most relevant to the regression problem. More specifically, PLS simultaneously decomposes both $\mathbf{X}$ and $\mathbf{Y}$ into N factors, stored in $\mathbf{T} \in \mathbb{R}^{M\times N}$ and $\mathbf{U} \in \mathbb{R}^{M\times N}$,

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathsf{T}} + \mathbf{E}, \quad \mathbf{P}^{\mathsf{T}}\mathbf{P} = \mathbf{I}_N,$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathsf{T}} + \mathbf{F}, \quad \mathbf{Q}^{\mathsf{T}}\mathbf{Q} = \mathbf{I}_N,$$

so that $\mathbf{t}_i^{\mathsf{T}}\mathbf{u}_i = \beta_i \to \max$, $i = 1, \ldots, N$. Then the solution to the regression problem is given by

$$\hat{\mathbf{Y}} = \hat{\mathbf{T}}\text{diag}(\boldsymbol{\beta})\mathbf{Q}^{\mathsf{T}} = \mathbf{X}\mathbf{W},$$

where $\hat{\mathbf{T}} = \mathbf{X}\mathbf{P}(\mathbf{P}^{\mathsf{T}}\mathbf{P})^{-1}$.

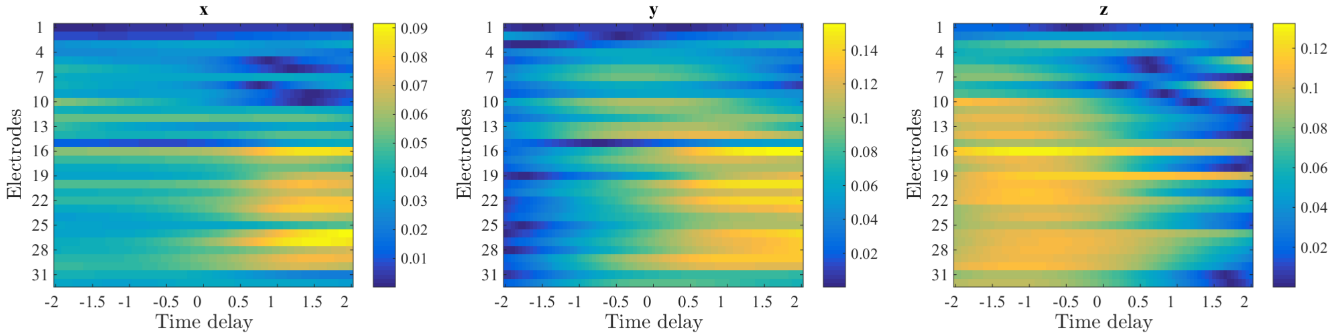
For multiway data we use NPLS, first proposed by [11]. The exact formulation of the algorithm can be found in [8].

3 Feature extraction for ECoG data

To test the proposed methods, we use feature extraction methods for ECoG-based classification and prediction of intended movements, most often reported successful in literature [12, 13, 14]. The feature description includes frequency- and time-domain features. Frequency-domain features are obtained with spectral transform (Short Time Fourier Transform or wavelet transform) or autoregressive analysis and represent time-dependent contributions of a range of frequencies into the signal. The time-domain features, referred to as LMP (local motor potentials) [12], are essentially low-passed ECoG time series $\mathbf{s}(t)$. Both time- and frequency-domain features are time delayed.

Time-domain features. The optimum latency value is chosen to maximize absolute linear cross-correlation between ECoG $\mathbf{s}(t)$ and target $\mathbf{y}(t)$ time series:

$$\tau_n^* = \underset{\tau\in[\tau_{\min},\tau_{\max}]}{\arg\max} \frac{|\sum_{i=1}^m s_n(t_i+\tau)y(t_i)|}{\sqrt{\sum_{i=1}^m s_n(t_i+\tau)s_n(t_i+\tau)}\sqrt{\sum_{i=1}^m y(t_i)y(t_i)}},$$



(a) x (b) y (c) z

Figure 1: Absolute values of cross-correlation between ECoG and target time series (left wrist) in time domain.

where $y(t)$ is the target time series for a given marker and dimension, and $s_n(t)$ is the ECoG time series for a given electrode.

As demonstrated by Fig. 1, the optimal latency τ^* might take both negative and positive values. Positive τ^* indicates that activity s_n that is most useful for prediction of the current position $y(t)$ is detected after that position was passed, which means that predictors based on such features are not causal.

After optimal latency τ^* depends the electrode position, but it turns out that τ^* are close for all electrodes when motion marker and dimension are fixed. Thus we select one value τ^* per marker-dimension pair.

Frequency-domain features. Each ECoG time series $\mathbf{s}_n(t)$, $n = 1, \dots, N_{\text{ch}}$ is transformed into frequency domain with wavelet transform. Here we use continuous wavelet (CWT) with Morlet as mother wavelet.

The voltage times series $\mathbf{s}(t)$ are sampled at 1000Hz for $N_{\text{ch}} = 32$ electrodes while the positions are sample at 120Hz. To convert the time series $\mathbf{s}(t)$, $\mathbf{y}(t)$ into the data sample $(\underline{\mathbf{D}}, \mathbf{Y})$, select M time points $t_1, \dots, t_M$ with time step δt.

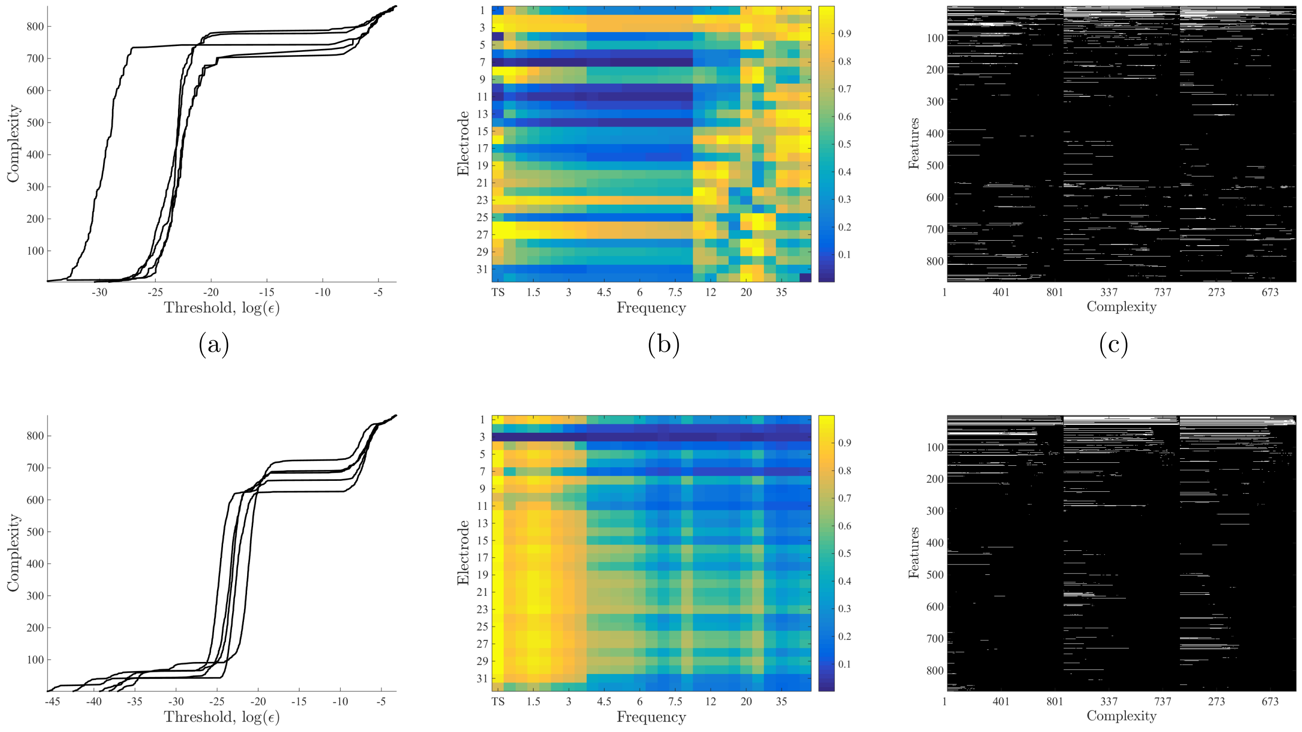
The feature matrix $\underline{\mathbf{X}}_m$ comprises information about the time series $\mathbf{s}(t)$ across the time period $t_m - \Delta < t \leq t_m$. The spatial component is represented by N_{ch} modes. To obtain $T \times F$ features in time-frequency domain, use the following procedure. Select F basic frequencies (scales) f_j, $j = 1, \dots, F$ and apply Morlet wavelet transform to all $s_n(t)$, $n = 1, \dots, N_{\text{ch}}$ at each center $t_1 \leq t_i \leq t_M$ and scale f_j, $j = 1, \dots, F$:

$$W_{ijn} = \frac{1}{\sqrt{|f_j|}} \sum_{t \leq t_M} \psi\left(\frac{t - t_i}{f_j}\right) s_n(t). \tag{9}$$

4 Experiments

In the computation experiments we used two feature extraction strategies, labeled 2D and 3D.

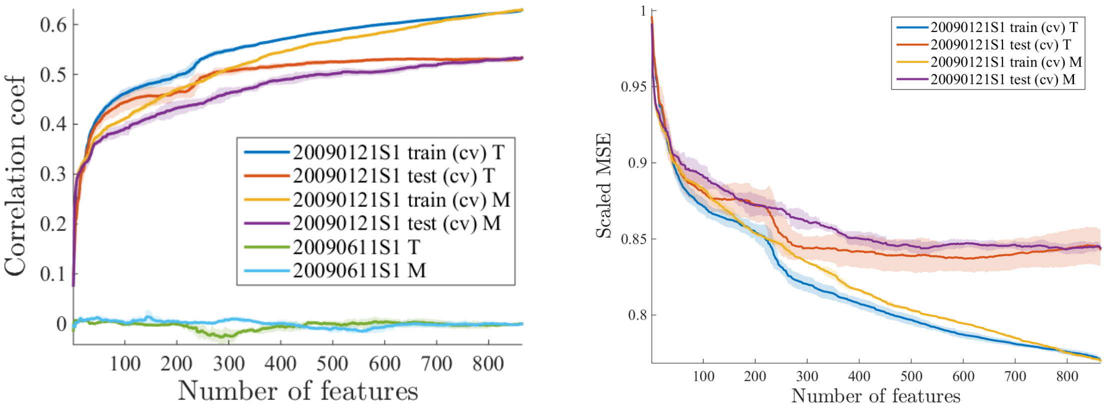
1. The 2D dataset includes the time-delayed ($\tau = 0.65s$) ECoG time series and wavelet



(a) (b) (c)

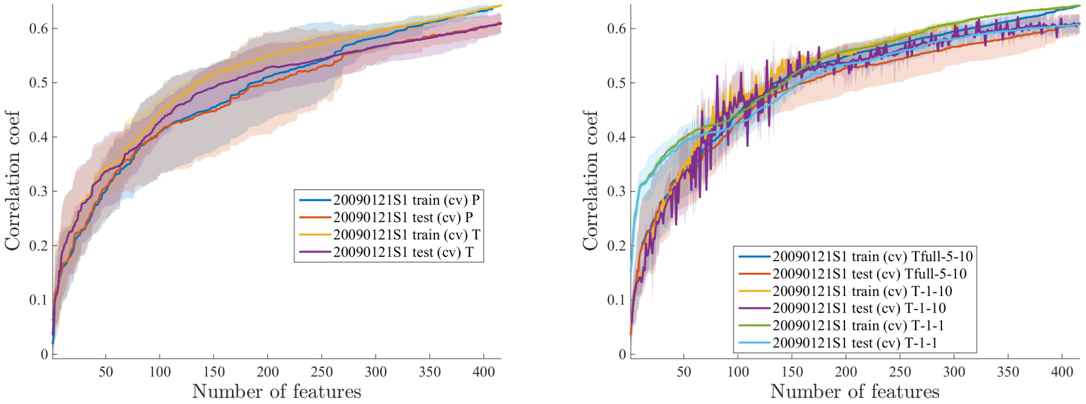
(d) (e) (f)

Figure 2: (a) Complexity by the threshold value ϵ. (b) Ratio of times $(i, j)-$th feature was selected into active feature set/ Averaged by threshold values and 5 cross-validation splits. (c) The diagram encodes inclusions of features into the model that were considered significant at $\alpha = 0.05$ level.



(a) Corr (b) sMSE

Figure 3: Forecasting quality by model complexity. Features are added by one in order defined by QPFS. The quality is measured as the correlation coefficient between forecasted projections of hand trajectory and the projections of the real trajectory (left wrist).



(a) PARAFAC vs Tucker

(b) Tucker

Figure 4: (a) Comparison of QPFS performance with similarity computed with PARAFAC or Tucker decomposition. (b) Results with Tucker decomposition, various parameters. Forecasting quality is measured as the correlation coefficient between forecasted projections of hand trajectory and the projections of the real trajectory (left wrist).

coefficients:

$$\underline{\mathbf{X}}_m \in \mathbb{R}^{F \times N_{\text{ch}}}, \quad \underline{\mathbf{X}}_{mjn} = \begin{cases} s_n(t_m + \tau), j = 1, \\ W_{mjn} \text{ for } j = 2, \dots, F+1, \end{cases} \quad n = 1, \dots, N_{\text{ch}}. \tag{10}$$

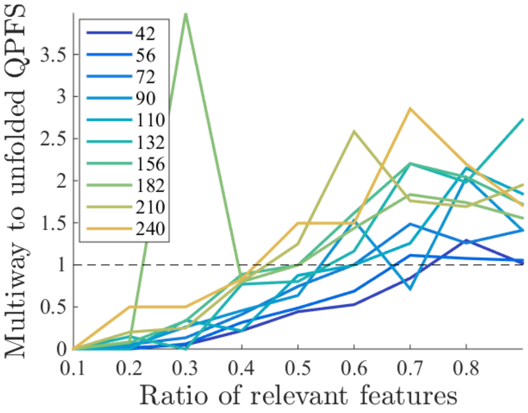
The time series were downsampled the data by the factor of 10. To create the data set we used the time step $\delta t = 0.05s$. We considered several frequency bands: 0.5–8Hz with 0.5Hz step, 9–18Hz with 3Hz and 20-45 with 5Hz step.

2. The 3D dataset contains three-way features with no time delay. 3D features explicitly include local history $\Delta_m = [t_m - \Delta t, t_m]$ of wavelet coefficients. To construct 3D dataset for $t_1, \dots, t_M$, select a finer grid of t_i, such that $|t_i \in \Delta_m| \geq T$, where T is the selected parameter, which controls how coarse is the summary of Δ_m. Split the time range Δ_m into T consecutive intervals δt_i, $i = 1, \dots, T$. For n-th electrode in $1, \dots, N$ the (i, j, n)-th element of three-way matrix $\underline{\mathbf{X}}_m \in \mathbb{R}^{T \times F \times N_{\text{ch}}}$ is given by averaging $W_{i'jn}$ over δ_i:

$$X_{mijn} = \frac{1}{|\delta t_i|} \sum_{i' \colon t_{i'} \in \delta t_i} W_{i'jn}. \tag{11}$$

Scalogram features were computed without downsampling with the following parameters: duration of local history time segment $\Delta t = 1s$ with step $\delta t = 0.05s$, $T = 20$, $F = 20$. The frequencies were chosen logarithmically spaced in the range $10 - 500$ Hz.

QPFS results. Figure 2 summarizes results of multi-way QPFS, applied to the 2D feature set (10). To evaluate performance of the QPFS algorithm, we splitted the part of the data set, correspondent to a time range from 5 to 645 seconds, into $K = 5$ folds to form a training set from four folds and a test set from one fold left. Each fold was used as test set once. The rest of the data (from 646 to 950) was used a sa hold-out set. Additionally we extracted the same



(a)

Figure 5: Comparison of selective abilities of multi-way and unfolded QPFS. The curves display ratio of F_1 measures (multi-way QPFS to unfolded QPFS) by the ratio of relevant features in the dataset. Each curve corresponds to a total number of features.

features from the time series measured during the experiment with the same monkey that took place six months later ("20090611").

The relaxed feature selection problem (7) was solved for each training set. The resulting structure variable $\underline{\mathbf{A}} \in [0,1]^{n_1 \times n_2}$ was then thresholded against some $\epsilon \in [0, 1)$ value to obtain an active feature set $\mathbf{X}_{\underline{\mathbf{A}}(\epsilon)}$. The quality of feature set $\underline{\mathbf{A}}(\epsilon)$ is evaluated as the forecasting quality $\mathcal{Q}(\mathbf{X}^{\text{test}}_{\underline{\mathbf{A}}(\epsilon)}\mathbf{w}_{\underline{\mathbf{A}}}, \mathbf{Y}^{\text{test}})$ of linear model (1), with parameters $\mathbf{w}_{\underline{\mathbf{A}}(\epsilon)}$ estimated at the training set.

Fig. 2(a) displays correlation coefficient between predicted $\hat{\mathbf{Y}}$ and true $\mathbf{Y}$ wrist trajectories against complexity $N(\epsilon) = \sum_{i,j} a_{ij}(\epsilon)$ of the model. The test quality stops increasing at about 300 features; hold-out quality stays approximately the same after about 100 features.

Fig. 2(b) shows how the complexity depends on the threshold value ϵ for each split.

Fig. 2(c) color-codes the ratio of times each feature was included into the active set among all ϵ values (averaged by cross-validation splits).

Fig. 2(d) demonstrates the results of statistical testing of model parameters $\mathbf{w}_{\underline{\mathbf{A}}(\epsilon)} \in \mathbb{R}^{N(\epsilon)\times 3}$ against zero. The diagram presents testing results at level $\alpha = 0.05$ for each dimension $d = 1, 2, 3$. Significant features χ_{ij} are colored white.

Comparison of multi-way to unfolded QPFS. Though the forecasting results produced by multi-way and unfolded QPFS are quite similar, Fig. 2 demonstrates that multi-way and unfolded QPFS tend to select different feature sets. To investigate the selective behaviour of QPFS and multi-way QPFS, we consider artificial multi-way dataset with structures and unstructured data. It is expected that the performance of multi-way QPFS applied to structured data should be better compared to the case of unstructured data. Also, when generating artificial data we control the level of multicollinearity between the features and the number of relevant features. Furthermore, in case of artificial data we know exactly which features χ_j are relevant: $j \in \underline{\mathbf{A}}$. This allows us to compute the F_1 measure

$$F_1 = \frac{2TP}{2TP + FP + FN}, \tag{12}$$

where TP and FP are the numbers of selected and left out relevant features:

$$TP = \left|\{j \in \underline{\mathbf{A}}\} \bigcap \{j \in \underline{\mathbf{A}}_\epsilon\}\right|, \quad FP = \left|\{j \not\in \underline{\mathbf{A}}\} \bigcap \{j \in \underline{\mathbf{A}}_\epsilon\}\right|,$$

and FN is the numbers of features, mistakenly attributed as irrelevant. The threshold value ϵ in (8) is selected so that the number of selected features is equal to the number of relevant features.

Comparison of QPFS and PLS. Figure 4 illustrates how correlation coefficient between predicted and true trajectories of the monkey's left wrist (contralateral to the electrodes placement) depend on number of features (or factors, in case of PLS) for matrix QP feature selection combined with linear regression and PLS regression. The data was split into train and test sets via 10-fold cross-validation strategy: the dataset is spit randomly into 10 sets and each set is used as the test set once. The quality presented on the is averaged over 10 splits. The models were learnt separately for different spatial dimensions x, y, z of the 3D target trajectory $\mathbf{Y}$. Each subfigure displays results for one of this dimensions.

The legend indicates date when the ECoG measurements underlying each curve were obtain. The bottom figures (figures 4(d–f)) demonstrate how the models learnt on the data taken on January 21, 2009 performed when applied to the data taken for the same monkey (S1) half a year later. As can be seen from figures 4(d–f), the decrease in quality is rather mild for both algorithms. Similar results were demonstrated in [15] and may be adherent to the data properties.

The figures imply that PLS provides more stable and more accurate trajectory reconstruction compared to QPFS.

Regularized unfolded PLS. Currently implemented: Sobolev regularization, smoothing-function regularization. Try Laplacian spatial smoothness constraint.

5 Conclusion

6 Appendix: derivation of NQPFS

Since $\underline{\mathbf{A}}$ is binary, an exact low-rank decomposition is possible:

$$\underline{\mathbf{A}} = \sum_{r=1}^{R} \mathbf{a}_1^{(r)} \circ \mathbf{a}_2^{(r)} \circ \mathbf{a}_3^{(r)}, \quad \mathbf{a}_1^{(r)} \in \mathbb{R}^{n_1},\ \mathbf{a}_2^{(r)} \in \mathbb{R}^{n_2},\ \mathbf{a}_3^{(r)} \in \mathbb{R}^{n_3}. \tag{13}$$

This allows to rewrite $\mathcal{Q}(\underline{\mathbf{A}})$ as

$$\begin{aligned}\mathcal{Q}(\underline{\mathbf{A}}) = \sum_{r=1}^{R} ||\mathbf{a}_2^{(r)}||_2^2 \cdot ||\mathbf{a}_3^{(r)}||_2^2 \cdot \mathbf{a}_1^{(r)\mathsf{T}} \mathbf{Q}_1 \mathbf{a}_1^{(r)} + ||\mathbf{a}_1^{(r)}||_2^2 \cdot ||\mathbf{a}_3^{(r)}||_2^2 \cdot \mathbf{a}_2^{(r)\mathsf{T}} \mathbf{Q}_2 \mathbf{a}_2^{(r)} + \\ ||\mathbf{a}_1^{(r)}||_2^2 \cdot ||\mathbf{a}_2^{(r)}||_2^2 \cdot \mathbf{a}_3^{(r)\mathsf{T}} \mathbf{Q}_3 \mathbf{a}_3^{(r)} + \underline{\mathbf{B}} \times_1 \mathbf{a}_1^{(r)} \times_2 \mathbf{a}_2^{(r)} \times_3 \mathbf{a}_3^{(r)}.\end{aligned} \tag{14}$$

With (14) the problem (7) solves iteratively, via alternate approach. At each step a quadratic programming with problem is solved. Let $\boldsymbol{\alpha}_i = [\mathbf{a}_i^{(1)\mathsf{T}}, \dots, \mathbf{a}_i^{(R)\mathsf{T}}]^\mathsf{T} \in \mathbb{R}^{nR}$ for $i = 1, 2, 3$ and $\boldsymbol{\alpha}^{(0)} = \mathbf{1}_{n_i R}$ be the initial approximation of $\boldsymbol{\alpha}_i$.

1. Solve the following problem with respect to $\boldsymbol{\alpha}_1$ with $\boldsymbol{\alpha}_2^{(k-1)}$, $\boldsymbol{\alpha}_3^{(k-1)}$ fixed:

$$\boldsymbol{\alpha}_1^{(k)} = \mathop{\arg\min}_{\boldsymbol{\alpha}\in\{0,1\}^{nR}} \boldsymbol{\alpha}_1^{\mathsf{T}}\left(\tilde{\mathbf{Q}}_1^{(k-1)}\boldsymbol{\alpha}_1 + \tilde{\mathbf{I}}_1^{(k-1)}\right) + \tilde{\mathbf{B}}_1^{(k-1)}\boldsymbol{\alpha}_1,$$

   where $\tilde{\mathbf{Q}}_1^{(k)}$ and $\tilde{\mathbf{I}}_1^{(k-1)}$ are block-diagonal with $r-$th blocks $\tilde{\mathbf{Q}}_1^{(k,r)}$ and $\tilde{\mathbf{I}}_1^{(k-1)}$:

$$\tilde{\mathbf{Q}}_1^{(k,r)} = ||\mathbf{a}_2^{(k,r)}||_2^2 \cdot ||\mathbf{a}_3^{(k,r)}||_2^2 \mathbf{Q}_1,$$

$$\tilde{\mathbf{I}}_1^{(k-1)} = (||\mathbf{a}_3^{(k,r)}||_2^2 \cdot \mathbf{a}_2^{(k,r)\mathsf{T}}\mathbf{Q}_2\mathbf{a}_2^{(k,r)} + ||\mathbf{a}_2^{(k,r)}||_2^2 \cdot \mathbf{a}_3^{(k,r)\mathsf{T}}\mathbf{Q}_3\mathbf{a}_3^{(r)})\mathbf{I}_{n_1},$$

   and

$$\tilde{\mathbf{B}}_1^{(k)} = [\tilde{\mathbf{B}}^{(k,1)\mathsf{T}}, \dots, \tilde{\mathbf{B}}^{(k,R)\mathsf{T}}]^{\mathsf{T}}, \quad \tilde{\mathbf{B}}^{(k,r)} = \underline{\mathbf{B}} \times_2 \mathbf{a}_2^{(k,r)} \times_3 \mathbf{a}_3^{(k,r)}.$$

2. Fix $\boldsymbol{\alpha}_1^{(k)}$, $\boldsymbol{\alpha}_3^{(k-1)}$, recompute $\tilde{\mathbf{Q}}_2^{(k)}$ and $\tilde{\mathbf{B}}_2^{(k)}$ and obtain $\boldsymbol{\alpha}_2^{(k)}$.

3. Fix $\boldsymbol{\alpha}_1^{(k)}$, $\boldsymbol{\alpha}_2^{(k)}$, recompute $\tilde{\mathbf{Q}}_3^{(k)}$ and $\tilde{\mathbf{B}}_3^{(k)}$ and obtain $\boldsymbol{\alpha}_3^{(k)}$.

The steps 1–3 repeat K times, which is, along with R, the parameter of NQPFS.

References

[1] Aleksandr Katrutsa and Vadim Strijov. Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. Expert Systems with Applications. (2017)

[2] Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. Brain computer interfaces, a review. Sensors (Basel), 12(2):1211–1279 (2012)

[3] S. Amiri, R. Fazel-Rezai, and V. Asadpour. A review of hybrid brain-computer interface systems. Advances in Human-Computer Interaction (2013).

[4] Andrey Eliseyev and Tetiana Aksenova. Penalized multi-way partial least squares for smooth trajectory decoding from lectrocorticographic (ecog). PLoS ONE, 11(5):e0154878 (2016)

[5] http://neurotycho.org/food-tracking-task.

[6] Qibin Zhao, Guoxu Zhou, Tülay Adali, and Andrzej Cichocki. Kernelization of tensorbased models for multiway data analysis. IEEE Signal Processing Magazine, 30(4):137–148 (2013)

[7] Yijun Wang, Shangkai Gao, and Xiaornog Gao. Common spatial pattern method for channel selection in motor imagery based bci. Engineering in Medicine and Biology 27th Annual Conference (2005)

[8] Andrey Eliseyev, Cecile Moro, Thomas Costecalde, Napoleon Torres, Sadok Gharbi, Corinne Mestais, Alim Louis Benabid, and Tatiana Aksenova. Iterative n-way partial least squares for a binary self-paced brain–computer interface in freely moving animals. J. Neural EngJournal of Neural Engineering, 8 (2011)

[9] C. Chen, D. Shin, Nakanishi-Y. Watanabe, H, and H. Kambara. Prediction of hand trajectory from electrocorticography signals in primary motor cortex. PLoS ONE 8(12): e83534, 2013.

[10] Andrzej Cichocki. Tensor decompositions: A new concept in brain data analysis? arXiv:1305.0395, 2013.

[11] R. Bro. Multiway calibration. multilinear pls. Chemometrics, 10(1):47–61, 1996.

[12] J. Kubánek, K. J. Miller, J. G. Ojemann, J. R. Wolpaw, and G. Schalk. Decoding flexion of individual fingers using electrocorticographic signals in humans. Journal Of Neural Engineering, 6(6):066001, 2009.

[13] Laurent Bougrain and Nanying Liang. Band-specific features improve finger flexion prediction from ecog. Jornadas Argentinas sobre Interfaces Cerebro Computadora – JAICC 2009, 2009.

[14] David T. Bundy, Mrinal Pahwa, Nicholas Szrama, and Eric C. Leuthardt. Decoding three-dimensional reaching movements using electrocorticographic signals in humans. Journal Of Neural Engineering, 13:026021, 2016.

[15] Z.C. Chao, Y. Nagasaka, and N. Fujii. Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkeys. Frontiers in Neuroengineering, 3:3, 2010.