Multi-way Feature Selection for ECoG-based Brain-Computer Interface

Anastasia Motrenko

Moscow Institute of Physics and Technology

Vadim Strijov

Dorodnicyn Computing Center of RAS

Abstract

The paper addresses the problem of designing Brain-Computer Interfaces. We solve 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 quadratic programming feature selection (QPFS), a recent approach to filtering-based feature selection proposed by Katrutsa and Strijov, "Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria". 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, hand movement prediction

1. Introduction

Brain-Computer Interface (BCI) system enhances its user's mental and physical abilities, providing a direct communication mean between the brain and a computer. BCIs aim at restoring damaged functionality of motorically or cognitively impaired patients. The problem of BCI design is far from being solved but this it promises great potential for assistive technology, rehabilitation devices and non-medical applications. In this paper we

Email address: anastasia.motrenko@phystech.edu (Anastasia Motrenko)

attempt to contribute to the BCI design, proposing a new method of feature selection in movement prediction and reconstruction.

Analysis of cortical activity during motor imagery is essential for BCI design. 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 [14, 1], we concentrate on the ElectroCorticoGraphic (ECoG) signals [10]. ECoG, as well as other invasive techniques, provides more stable recordings and better resolution in temporal and spatial domains than its non-invasive counterparts.

The first step to predicting intended movements is learning to reconstruct actual movements from cortical activity. We address the problem of continuous hand trajectory reconstruction. The subdural ECoG signals are measured across 32 or 64 channels as the subject is moving its hand [18]. Once the ECoG signals are transformed into informative features, the problem of trajectory reconstruction is a regression problem. Feature extraction involves application of some spectro-temporal transform to the ECoG signals from each channel [13, 15, 9]. Since the resulting spatial-temporal-spectral representation is highly redundant, various feature selection and dimensionality reduction techniques are used [17, 16] to extract only the most relevant features.

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. Unfolding multi-way data into flat matrices might lead to neglecting important dependencies present in the unfolded dimension of the multi-way data. Contrarily, multi-way approaches preserve the data structure and improve regression quality, as was demonstrated by [17] for the Partial Least Squares (PLS) regression. PLS regression and its extensions for multi-way data [10, 17] have proven their efficiency in ECoG-based hand trajectory reconstruction [17, 9, 7]. Similarly to the original PLS relying on the Singular Value Decomposition, multi-way extensions of PLS rely on multi-way decompositions, such as Tucker decomposition or PARAFAC [8]. Additinally, several regularisation techniques were proposed to increase its stability [10] and reduce overfitting.

In this paper we consider an alternative feature selection method for multi-way data. We propose a multi-way formulation of a recent approach to filtering feature selection by Katrutsa [11], Quadratic Programming Feature Selection (QPFS). Filtering methods separate feature selection from regression model training. This makes them computationally efficient even for large dimensionalities. A special feature of QPFS is the ability consider the relationships between features. QPFS is formulated as a quadratic program which minimizes correlation between features while maximizing feature relevance.

The original QPFS ignores multi-way structure of the data. To adapt this powerful method to feature selection in movement prediction for BCI construction, we propose a multi-way extension of QPFS. We introduce a separate feature similarity matrix for each modality of the feature description. This reduces dimensionality of optimization problem, which makes the proposed approach applicable even in case of high dimensionality. We compare the original and multi-way QPFS applied to trajectory reconstruction problem and show that proposed modification without loss in quality. We also compare both versions of QPFS and PLS regression.

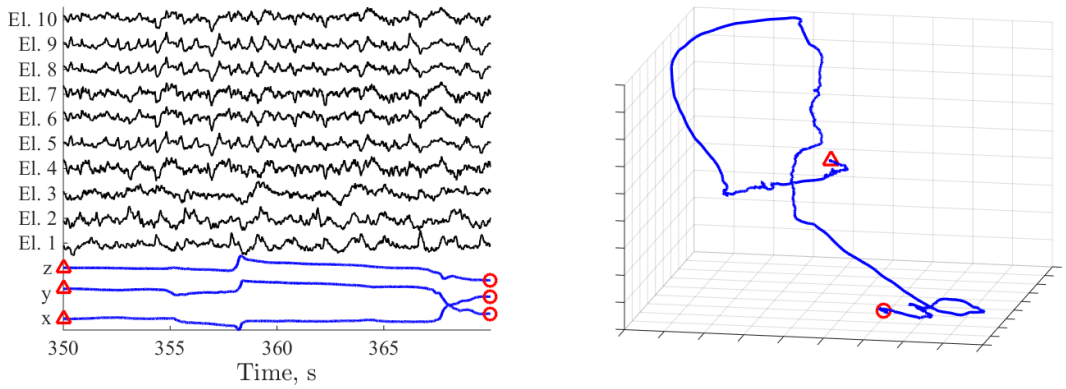


Figure 1: Left: Extracts (350–370s) from voltage and wrist position time series for monkey A. Right: 3D wrist trajectory for the same extract.

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 $1, \ldots, N_{\text{ch}}$, and multivariate target time series $\mathbf{y}(t) \in \mathbb{R}^3$ with 3D wrist[1] coordinates. These time series are converted to the data sample $(\underline{\mathbf{D}}, \mathbf{Y})$:

$$\underline{\mathbf{D}} \in \mathbb{R}^{T \times F \times N_{\text{ch}} \times M}, \; D_{(m,:,:,:)} = \underline{\mathbf{X}}_m, \quad \mathbf{Y} = [\mathbf{y}_1^\mathsf{T}, \ldots, \mathbf{y}_M^\mathsf{T}]^\mathsf{T}, \tag{1}$$

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 $[s_n(t_m - \Delta t), \ldots, s_n(t))]$ along the channel n, $n = 1, \ldots, N_{\text{ch}}$. 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 hand trajectory $\mathbf{Y}$ given $\underline{\mathbf{X}}_m$, $m = 1, \ldots M$. The reconstructed trajectory $\hat{\mathbf{Y}}$ approximates the real one as a linear combination of features:

$$\hat{\mathbf{y}}_m = \text{vec}\big(\underline{\mathbf{X}}_m\big)^\mathsf{T} \hat{\mathbf{w}}, \tag{2}$$

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:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} ||\hat{\mathbf{Y}} - \mathbf{Y}||_2^2. \tag{3}$$

Feature selection.. Due to the fact that ECoG measurements are highly correlated both in time and space, the problem (2), (3) is instable. To decrease computational cost and increase stability of reconstruction we use feature selection methods and add regularization term to the problem (3).

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

$$\mathbf{X} = [\text{vec}(\underline{\mathbf{X}}_1), \ldots, \text{vec}(\underline{\mathbf{X}}_M)]^\mathsf{T} = \tag{4}$$

[1]The dataset [18] includes trajectories of elbows, shoulders and possibly others. In the experiments we used wrist positions of the hand contralateral to the electrodes placements as targets.

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

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 (3) for $\mathbf{X}_{\underline{\mathbf{A}}}$.

To evaluate forecasting quality, we use scaled MSE

$$\text{sMSE}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{\sum_{m=1}^{M} ||\hat{\mathbf{y}}_m - \mathbf{y}_m||_2}{\sum_{m=1}^{M} ||\bar{\mathbf{y}} - \mathbf{y}_m||_2}, \quad \bar{\mathbf{y}} = \frac{1}{M} \sum_{i=1}^{M} y_m, \tag{5}$$

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

$$\text{corr}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{\text{cov}(\hat{\mathbf{y}}, \mathbf{y})}{\sqrt{\text{cov}(\hat{\mathbf{y}}, \hat{\mathbf{y}}) \text{cov}(\mathbf{y}, \mathbf{y})}}. \tag{6}$$

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 [11]. 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 [11] 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

$$\mathbf{a} = \underset{\mathbf{a} \in \{0,1\}^N}{\arg\min} \left(\mathbf{a}^{\intercal} \mathbf{Q} \mathbf{a} - \mathbf{b}^{\intercal} \mathbf{a}\right), \tag{7}$$

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

$$q_{ij} = |\text{corr}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)| \approx \frac{1}{M} \left[\mathbf{X}^{\intercal} \mathbf{X}\right]_{ij}. \tag{8}$$

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 $\mathbf{Y}$:

$$b_i = \frac{1}{3} \sum_{n=1}^{3} |\text{corr}(\boldsymbol{\chi}_i, \mathbf{y}_n)|. \tag{9}$$

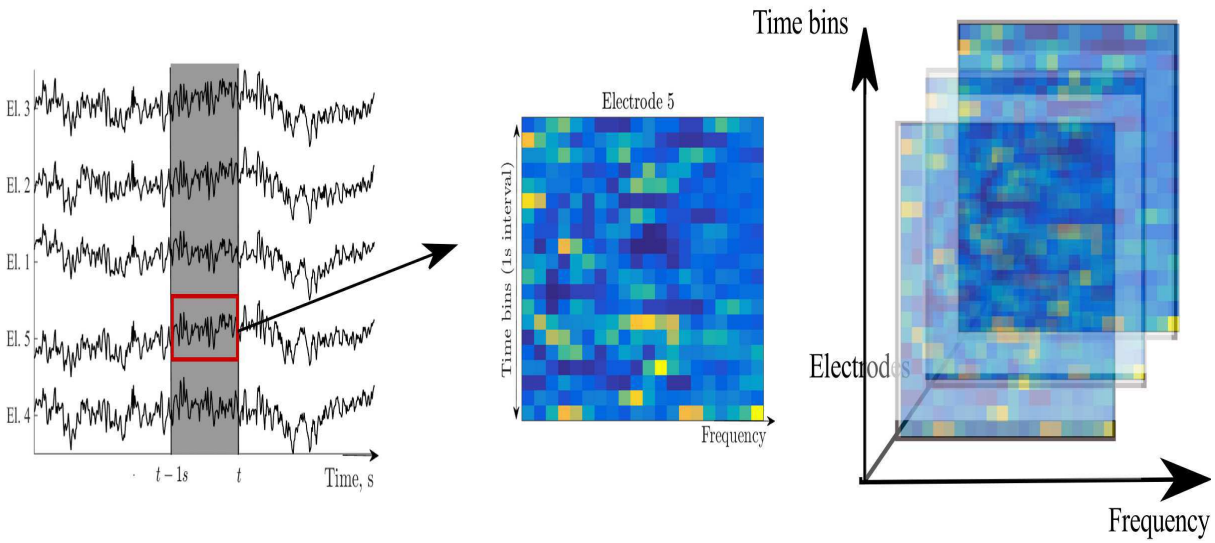


Figure 2: Example of feature construction procedure. For each electrode a one-second long historical interval $[t_m - \Delta t, t_m]$ undergoes wavelet transformation (13) and thus obtains feature description in spectral-temporal domain. Merging spectral-temporal feature matrices for all electrodes, one obtains 3D feature description $\underline{\mathbf{X}}_m$ for the time point t_m.

Note that there are other ways to define $\mathbf{Q}$ and $\mathbf{b}$ besides the correlation coefficient. For example, [11] also considers mutual information and normalized feature significance as similarity and relevance measures.

The problem (7) 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:

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

Multi-way QPFS.. The problem (7) is formulated for the two-way data. In case of four-way[2] 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, since instead of feature vector $\boldsymbol{\chi}_i \in \mathbb{R}^n$ we now have $\boldsymbol{\chi}_{ijk} = \underline{\mathbf{D}}_{(:,i,j,k)} \in \mathbb{R}^{n_1 n_2 n_3}$. The obvious solution to this would be to flatten the features $\underline{\mathbf{D}}$ by vectorizing (4) each d-mode matrix $\underline{\mathbf{X}}_m$ and then proceed with the original QPFS (7). The weak spot of such approach is computing similarity matrix $\mathbf{Q}$. Since $\mathbf{X}$ contain multiple correlated features, $\mathbf{Q}$ becomes closer to singular as the number of features grows, even if it is positive-definite by definition (8). The construction of optimization problem becomes the most difficult part of QPFS. To overcome this problem, we incorporate the multi-way structure of ECoG features $\underline{\mathbf{X}}_m$ and propose a multi-way formulation of QPFS. More specifically, we assign one similarity matrix for each mode: $\mathbf{Q}_1 \in \mathbb{R}^{n_1 \times n_1}$, $\mathbf{Q}_2 \in \mathbb{R}^{n_2 \times n_2}$, $\mathbf{Q}_3 \in \mathbb{R}^{n_3 \times n_3}$. The relevance matrix is the same size as $\underline{\mathbf{X}}_m$.

We use the following notation. 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 (7) reformulates as follows:

$$\underline{\mathbf{A}} = \underset{\underline{\mathbf{A}} \in \{0,1\}^{n_1 \times n_2 \times n_3}}{\arg\min} \left(\sum_{d=1}^{3} (\underline{\mathbf{A}} \times_1 \mathbf{Q}_d) * \underline{\mathbf{A}} - \underline{\mathbf{B}} * \underline{\mathbf{A}} \right) \times_1 \mathbf{1}_{n_1} \times_2 \mathbf{1}_{n_2} \times_3 \mathbf{1}_{n_3}, \tag{10}$$

[2]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 $\underline{\mathbf{X}}_m \in \mathbb{R}^{n_1 \times \cdots \times n_d}, d \geq 2$.

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 (10) is based on low rank decomposition of $\underline{\mathbf{A}}$, which allows to solve the problem (10) 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:

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

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

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

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

The relevance definition (9) generalizes straightforwardly to the three-way case:

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

Linear relaxation of (10).. The problem (10) is the integer optimization problem, which is not convex. To allow for more efficient solution, we have to relax non-convex constraint $\underline{\mathbf{A}} \in \{0, 1\}^{n_1 \times n_2 \times n_3}$ into $\hat{\underline{\mathbf{A}}} \in [0, 1]^{n_1 \times n_2 \times n_3}$. After the solution $\hat{\underline{\mathbf{A}}}$ of the relaxed problem is found, we threshold $\hat{\underline{\mathbf{A}}}$ to $\{0, 1]\}$

$$\underline{\mathbf{A}}(\epsilon) = [a_{ijk}], \quad a_{ijk} = \begin{cases} 1 \text{ if } \hat{a}_{ijk} \geq \epsilon, \\ 0 \text{ otherwise} \end{cases} \tag{11}$$

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

$$\boldsymbol{\chi}_{ijk} \preceq \boldsymbol{\chi}_{i'j'k'} \Leftrightarrow \hat{a}_{ijk} \leq \hat{a}_{i'j'k'}. \tag{12}$$

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 $\mathbf{X}$ (4) and $\mathbf{Y}$ (1) into N factors, stored in $\mathbf{T} \in \mathbb{R}^{M\times N}$ and $\mathbf{U} \in \mathbb{R}^{M\times N}$,

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

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

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

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

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

For multiway data we use NPLS, first proposed by [4]. The exact formulation of the algorithm can be found in [9].

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, 3, 5]. The feature description includes frequency- and time-domain features. Frequency-domain features are obtained with spectral transform (we use wavelet transform, but other options, such as short time Fourier transform or autoregressive analysis are possible) 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.

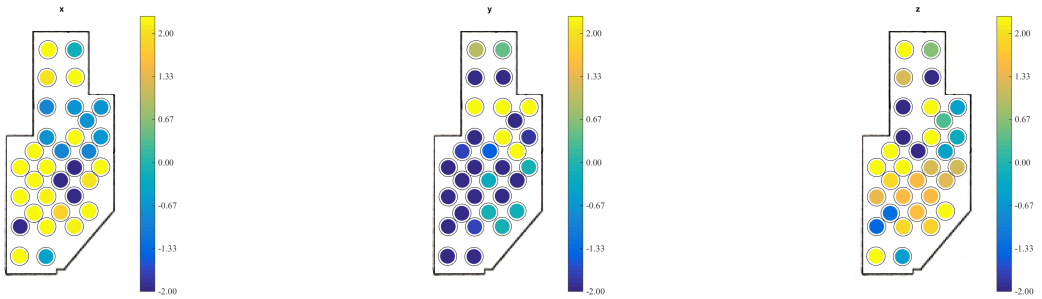
Time-domain features.. The optimum latency value is chosen to maximize absolute linear cross-correlation between ECoG $\mathbf{s}(t)$ and target $\mathbf{y}(t)$ time series:

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

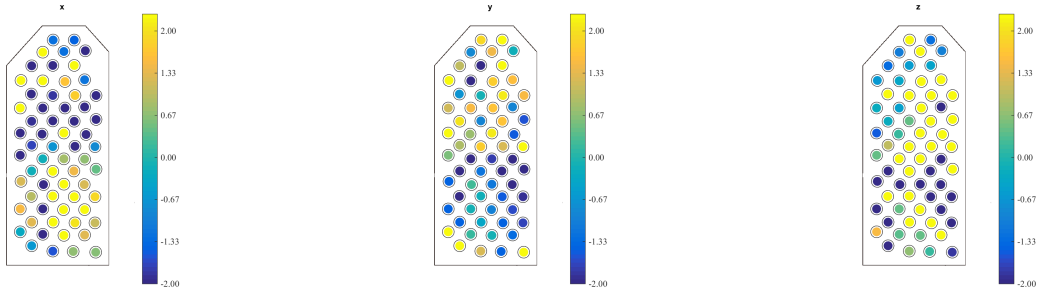
where $y(t)$ is the target time series for a chosen marker and dimension, and $s_n(t)$ is the ECoG time series for a given electrode.

As demonstrated by Fig. 3, 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.

Optimal latency τ^* values depend both on the electrode position and the spatial pattern of this dependency varies between x, y, z dimensions of target time series and subjects. Hence we select zero latency for all subjects.



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



(d) x (e) y (f) z

Figure 3: Absolute values of cross-correlation between ECoG and target time series (wrist positions) in time domain for monkeys A (top, 32 electrodes) and K1 (bottom, 64 electrodes).

Frequency-domain features.. The voltage times series $\mathbf{s}(t)$ are sampled at 1000Hz for $N_{\text{ch}} = 32$ or $N_{\text{ch}} = 64$ 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 $t_1, \ldots, t_M$ with time step δt.

The feature matrix $\underline{\mathbf{X}}_m$ comprises spatial, temporal and spectral information about the time series $\mathbf{s}(t)$ across the time period $[t_m - \Delta, t_m]$. Fig. 2 illustrates the process of feature extraction. The spatial component is represented by N_{ch} electrodes. Each ECoG time series $\mathbf{s}_n(t)$, $n = 1, \ldots, N_{\text{ch}}$ is transformed into frequency domain with wavelet transform. Here we use continuous wavelet (CWT) with Morlet as mother wavelet. To obtain $T \times F$ features in time-frequency domain, use the following procedure. Select F basic frequencies (scales) f_j, $j = 1, \ldots, F$ and apply Morlet wavelet transform to all $s_n(t)$, $n = 1, \ldots, N_{\text{ch}}$ at each center $t_1 \le t_i \le t_M$ and scale f_j, $j = 1, \ldots, F$:

$$W_{ijn} = \frac{1}{\sqrt{|f_j|}} \sum_{t \le t_M} \psi\left(\frac{t - t_i}{f_j}\right) s_n(t). \tag{13}$$

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 coefficients:

$$\underline{\mathbf{X}}_m \in \mathbb{R}^{F \times N_{\text{ch}}}, \quad \underline{\mathbf{X}}_{mjn} = \begin{cases} s_n(t_m + \tau), j = 1, \\ W_{mjn} \text{ for } j = 2, \ldots, F+1, \end{cases} \quad n = 1, \ldots, N_{\text{ch}}. \tag{14}$$

   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 $t_1, \ldots, t_M$, select a finer grid of t_i, such that $|t_i \in \Delta_m| \ge T$, 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, $i = 1, \ldots, T$. For n-th electrode in $1, \ldots, N$ 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:

$$X_{mijn} = \frac{1}{|\delta t_i|} \sum_{i' \colon t_{i'} \in \delta t_i} W_{i'jn}. \tag{15}$$

   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.

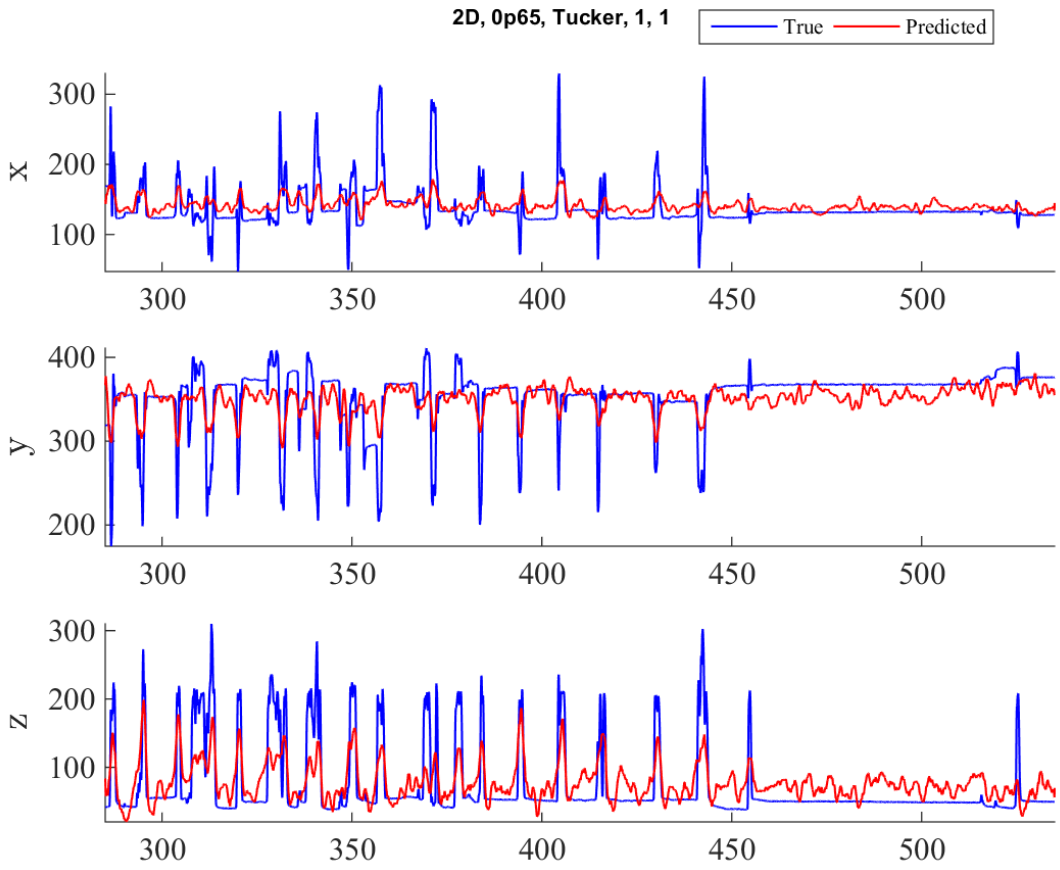
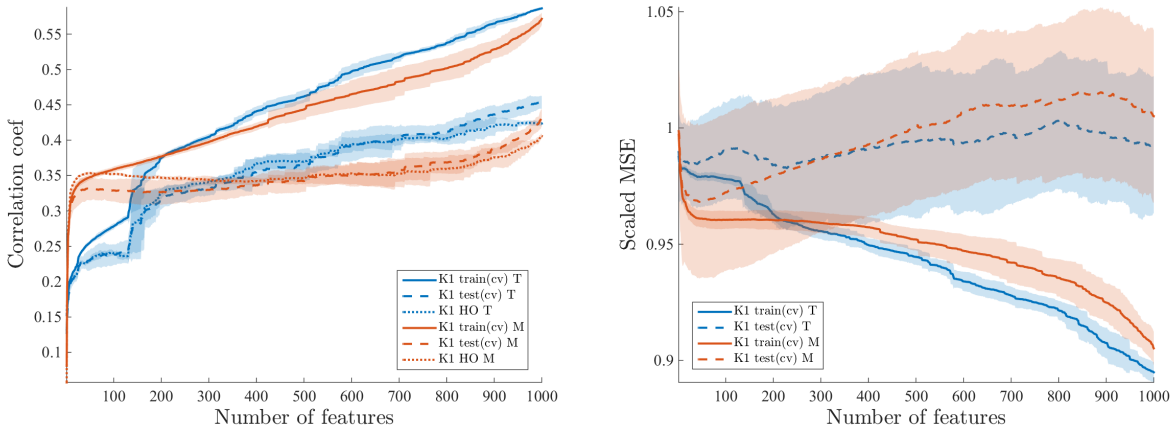
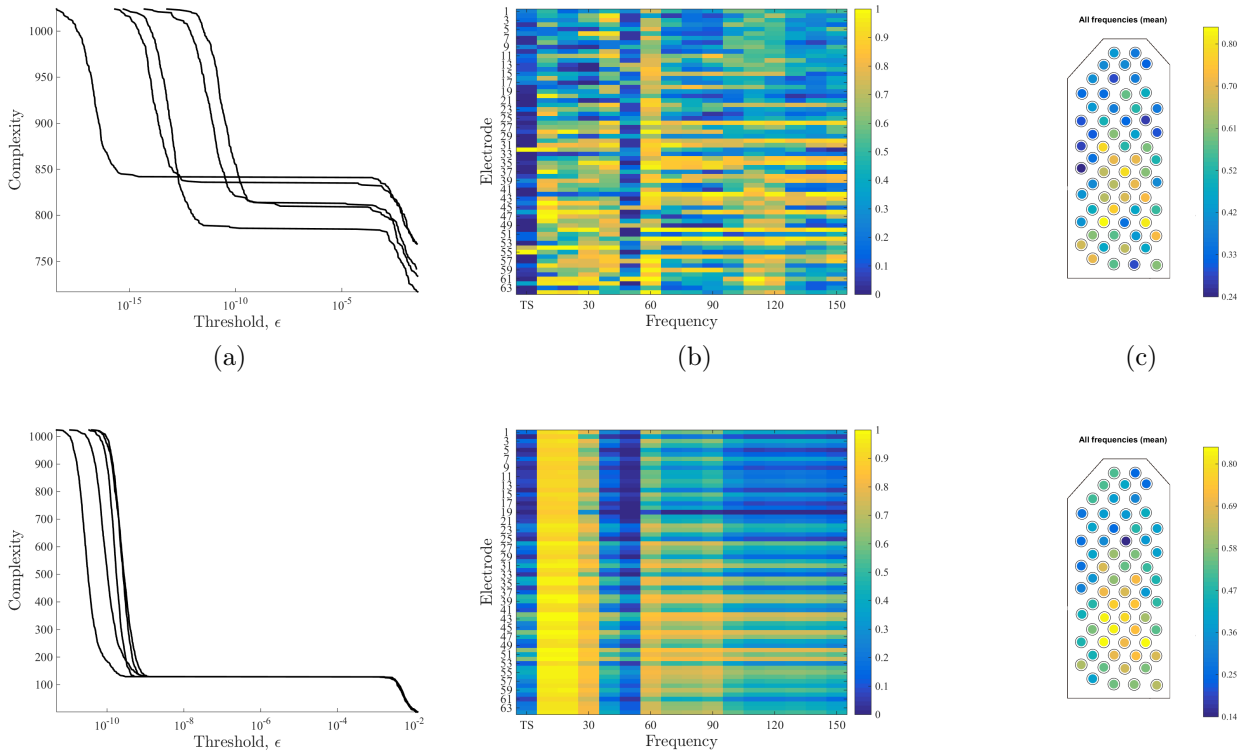


Figure 4: Segment of the forecasted time series. Linear regression, 50 best features according to multi-way QPFS.



(a) Corr (b) sMSE

Figure 5: Forecasting quality by model complexity. Features are added by one in order (12) 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) (b) (c)

(d) (e) (f)

Figure 6: (a) Complexity by the threshold value ϵ. (b) Evaluation of electrode-frequency pairs importance. Importance is measured as feature rank (12), averaged over cross-validation splits. (c) Electrode ranks, averaged over frequencies.

QPFS results.. In this section we compare performance of original QPFS and multi-way QPFS[3]. Figure 6 summarizes results of multi-way QPFS, applied to the 2D feature set (14). 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 as a hold-out set.

The relaxed feature selection problem (10) was solved for each training set. The resulting structure variable defined some ranking of features (12). We say that $\chi_{(i,j)}$ feature is ranked $n-$th, if it better than $n-1$ features. Since higher ranked features are more likely to be included into the model, we measured feature importance as its rank, averaged over cross-validation splits.

To obtain to obtain an active feature set $\mathbf{X}_{\underline{\mathbf{A}}(\epsilon)}$ we and thresholded $\hat{\underline{\mathbf{A}}}$ against some $\epsilon \in [0, 1)$ value. 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 (2), with parameters $\mathbf{w}_{\underline{\mathbf{A}}(\epsilon)}$ estimated at the training set.

Fig. 4, 5, and 6 exemplify QPFS results for 2D feature set. Fig. 4 demonstrates an example of forecast, obtained a monkey's hand trajectory with 25 features, selected by multi-way QPFS from 2D feature set. As can be seen, the reconstructed trajectory more or less follows the peaks in the original trajectory. However, it is too jerky in the "still" regions (later than 450 ms) which may cause disturbances for the BCI user. Perhaps a mixture model, which operates several different models (say, one for rigorous movement and one for stillness) to obtain the final forecast, would do better in this case. We leave more complex modelling technique as well as postprocessing out of the scope of this paper, since our goal is to propose a feature selection method.

Fig. 5 shows quality curves for unfolded and multi-way QPFS. Fig. 5(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. 5(a) displays scaled MSE of the predicted $\hat{\mathbf{Y}}$ trajectory against complexity.

Fig. 6(a) shows how the complexity N depends on the threshold value ϵ for each split. QPFS seems to partition the feature set into several groups with approximately the same value of structure variable $\hat{\underline{\mathbf{A}}}$. Fig. 6(b) color-codes importances of each feature (electrode-frequency pair). The first column of the color-coded matrix corresponds to decimated ECoG time series. Fig. 6(c) shows color-coded importances of each electrode, averaged by frequencies. The electrodes are positioned in accordance with the ECoG electrode placement used in the experiments (monkey A). Though the forecasting results 5 produced by multi-way and unfolded QPFS are quite similar, Fig. 6 demonstrates that multi-way and unfolded QPFS tend to select different feature sets.

In case of 3D feature set we had to additionally split the data into batches since due to the large dimension of the feature set. Fig. 7 shows how quality curves change their shape

[3]The code for computational experiments is available at https://github.com/Anastasia874/ECoGFeatureSelection

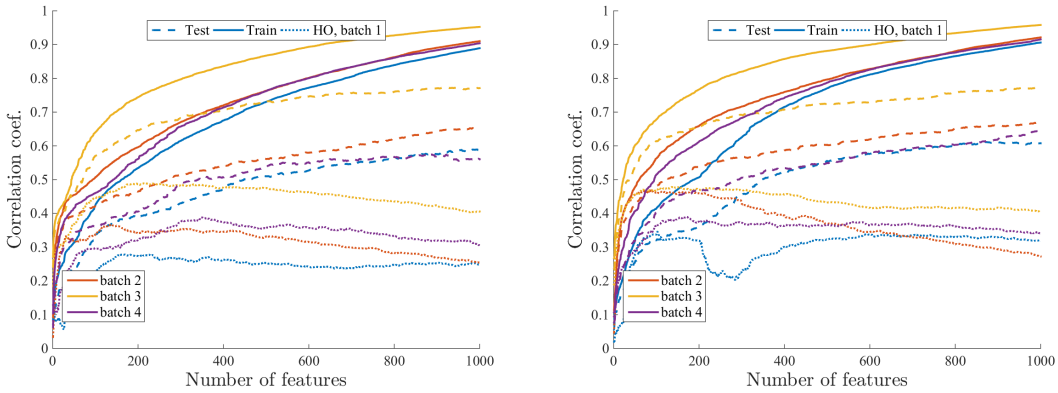


Figure 7: Forecasting quality measures as correlation coefficient between the original wrist trajectory and the reconstructed trajectory. Left: unfolded QPFS, right: multi-way QPFS.

from batch to batch. Fig. 8 shows how distribution of feature importances in electrode-frequency domain changes from batch to batch. It can be seen, for example, that unfolded QPFS places more emphasis on the low frequencies and anterior electrodes compared to multi-way QPFS. We selected the 10-th time bin for demonstration, but other bins display similar patterns. The right part of each subfigure of Fig. 8 shows the spatial distribution of importances averaged over all frequencies for the 10-th time bin.

Comparison of QPFS and PLS.. We compared the proposed algorithm to unfolded PLS and NPLS, the way it was formulated in [9]. Table 1 compares unfolded QPFS, multi-way QPFS (label NQPFS by analogy with NPLS), PLS and NPLS in terms of correlation coefficient (6). Each algorithm was allowed to select $N = 10, 25, 50$ or 200 features (or components, in case of PLS and NPLS). For QPFS and NQPFS we then estimated parameters of linear model (2) with no additional regularization. In case of PLS and NPLS, parameters come as the part of solution to the component construction problem. For each data set (5 data sets for monkey A and 3 data sets for monkey K) and each N the best result is given in bold.

In addition to previously mentioned correlation coefficient (6) and scaled MSE (5) we used dynamic time warping (DTW) [Strijov, Goncharov?] distance and mean absolute difference error (MADE) between $\hat{\mathbf{Y}}$ and $\mathbf{Y}$. DTW is used as distance measure when the compared sequences must be alighted before comparison and represents the cost of best alignment. MADE measures smoothness of the reconstructed trajectory as

$$\text{MADE}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{\sum_{m=1}^{M} |\hat{\mathbf{y}}'_m - \mathbf{y}'_m|}{\sum_{m=1}^{M} |\bar{\mathbf{y}}' - \mathbf{y}'_m|}.$$

This metric is important in trajectory reconstruction in the context of BCI design. Results of comparison by all four metrics are summarised in Fig. 9. Fig. 9(a) displays average values of these criteria for $N = 10, 25, 50, 100, 200, 500$. For each algorithm we calculated how many times it placed first (rank one), second (rank two) and so on, and averaged their

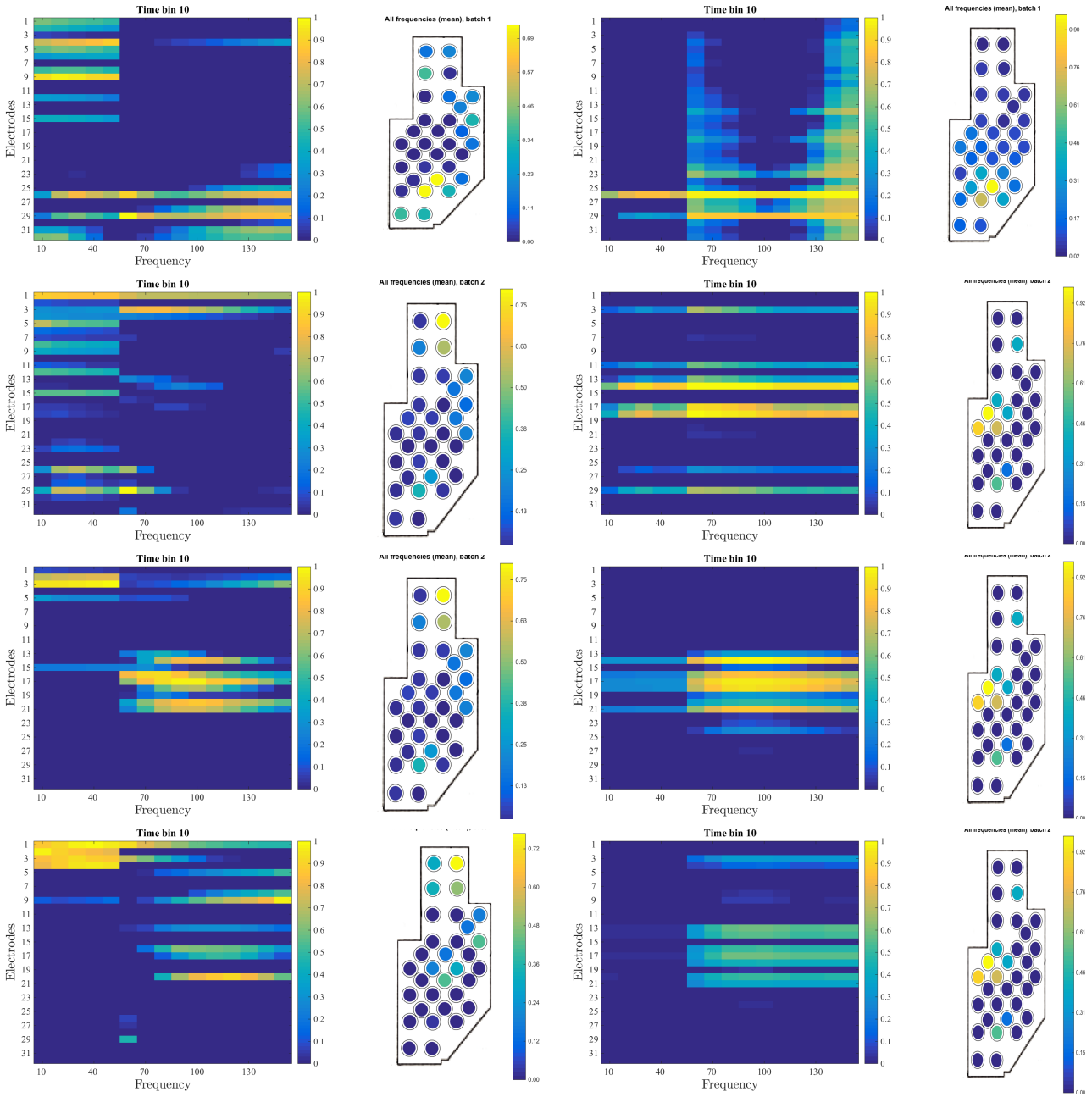
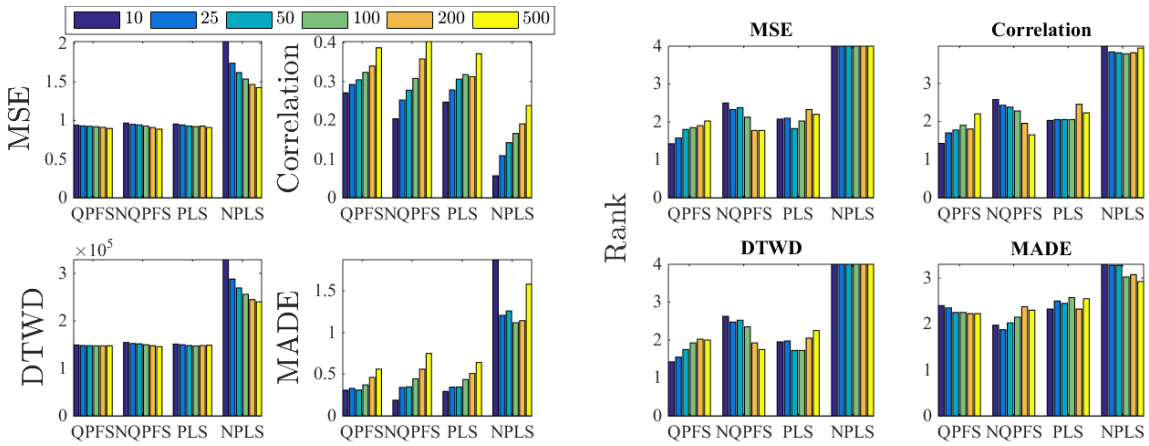


Figure 8: Selection frequency of electrode-frequency pairs, time bin 10; left: unfolded QPFS, right: multi-way QPFS.

Table 1: All monkeys, correlation coefficient.

Monkey, date	Algorithm	$N = 10$	$N = 25$	$N = 200$	$N = 500$
A, 20090116	QPFS	0.285 ± 0.016	0.295 ± 0.017	0.342 ± 0.007	0.454 ± 0.025
	NQPFS	$\mathbf{0.287 \pm 0.068}$	$\mathbf{0.384 \pm 0.029}$	$\mathbf{0.481 \pm 0.006}$	$\mathbf{0.482 \pm 0.008}$
	PLS	0.255 ± 0.011	0.29 ± 0.013	0.33 ± 0.008	0.481 ± 0.011
	NPLS	0.0477 ± 0.007	0.1 ± 0.006	0.196 ± 0.017	0.312 ± 0.060
K, 20090525	QPFS	$\mathbf{0.214 \pm 0.087}$	$\mathbf{0.237 \pm 0.088}$	$\mathbf{0.358 \pm 0.144}$	$\mathbf{0.384 \pm 0.155}$
	NQPFS	0.136 ± 0.036	0.15 ± 0.033	0.276 ± 0.102	0.35 ± 0.125
	PLS	0.198 ± 0.043	0.221 ± 0.091	0.339 ± 0.124	0.309 ± 0.113
	NPLS	0.042 ± 0.014	0.0837 ± 0.023	0.245 ± 0.081	0.259 ± 0.092
K, 20090527	QPFS	$\mathbf{0.313 \pm 0.022}$	$\mathbf{0.337 \pm 0.017}$	$\mathbf{0.362 \pm 0.023}$	$\mathbf{0.371 \pm 0.019}$
	NQPFS	0.165 ± 0.011	0.165 ± 0.008	0.243 ± 0.014	0.306 ± 0.009
	PLS	0.237 ± 0.007	0.283 ± 0.008	0.319 ± 0.018	0.313 ± 0.014
	NPLS	0.0901 ± 0.010	0.2 ± 0.010	0.284 ± 0.013	0.286 ± 0.009
K, 20090602	QPFS	$\mathbf{0.33 \pm 0.006}$	$\mathbf{0.351 \pm 0.008}$	0.374 ± 0.006	0.38 ± 0.009
	NQPFS	0.265 ± 0.006	0.277 ± 0.009	$\mathbf{0.387 \pm 0.008}$	$\mathbf{0.536 \pm 0.006}$
	PLS	0.301 ± 0.009	0.322 ± 0.007	0.355 ± 0.011	0.328 ± 0.024
	NPLS	0.0851 ± 0.005	0.13 ± 0.009	0.242 ± 0.006	0.252 ± 0.028
A, 20081127	QPFS	$\mathbf{0.294 \pm 0.028}$	0.335 ± 0.018	0.39 ± 0.019	0.458 ± 0.037
	NQPFS	0.218 ± 0.010	$\mathbf{0.342 \pm 0.016}$	$\mathbf{0.474 \pm 0.011}$	0.465 ± 0.045
	PLS	0.27 ± 0.008	0.321 ± 0.013	0.368 ± 0.021	$\mathbf{0.466 \pm 0.043}$
	NPLS	0.0426 ± 0.006	0.0743 ± 0.010	0.161 ± 0.005	0.231 ± 0.021
A, 20081224	QPFS	$\mathbf{0.257 \pm 0.060}$	0.271 ± 0.061	0.305 ± 0.065	0.404 ± 0.089
	NQPFS	0.168 ± 0.008	0.192 ± 0.004	$\mathbf{0.369 \pm 0.137}$	$\mathbf{0.422 \pm 0.092}$
	PLS	0.253 ± 0.009	$\mathbf{0.279 \pm 0.009}$	0.273 ± 0.082	0.421 ± 0.087
	NPLS	0.0495 ± 0.012	0.082 ± 0.004	0.104 ± 0.043	0.19 ± 0.076
A, 20090121	QPFS	$\mathbf{0.27 \pm 0.007}$	0.276 ± 0.007	0.322 ± 0.013	0.368 ± 0.013
	NQPFS	0.254 ± 0.013	$\mathbf{0.337 \pm 0.005}$	$\mathbf{0.394 \pm 0.005}$	$\mathbf{0.389 \pm 0.009}$
	PLS	0.225 ± 0.003	0.244 ± 0.019	0.242 ± 0.019	0.383 ± 0.014
	NPLS	0.0426 ± 0.008	0.0695 ± 0.009	0.0992 ± 0.010	0.175 ± 0.025
A, 20090611	QPFS	0.201 ± 0.009	0.233 ± 0.015	0.264 ± 0.019	$\mathbf{0.276 \pm 0.014}$
	NQPFS	0.136 ± 0.015	0.169 ± 0.010	0.24 ± 0.007	0.273 ± 0.003
	PLS	$\mathbf{0.234 \pm 0.008}$	$\mathbf{0.263 \pm 0.014}$	$\mathbf{0.274 \pm 0.005}$	0.271 ± 0.007
	NPLS	0.0533 ± 0.007	0.127 ± 0.008	0.193 ± 0.006	0.196 ± 0.008



(a) (b)

Figure 9: Left: absolute values of cross-correlation between ECoG and target time series (wrist position) in time domain.

rankings over 8 data sets. Fig. 9(b) reports average rankings of the algorithms. Here less is better.

5. Conclusion

The paper proposes a multi-way formulation of the quadratic programming feature selection approach. QPFS is a flexible and efficient approach, which allows to select select most relevant features from a highly multi-correlated set. Our modification is designed for multi-way structured data. We exploit the data structure to formulate a version of QPFS suitable even for high dimensions. The proposed modification of QPFS is applied to the problem of hand trajectory prediction. We observed that multi-way QPFS provided similar results to those of QPFS in terms of regression quality. Also, the quality of simple linear regression based on QPFS-selected features is comparable with the quality PLS regression.

Future research.. Here we have adopted the traditional approach to brain signal analysis based on spectral transforms (continuous wavelet transform in this case). However, many powerful algorithms [2] were proven able to compete with feature extraction based on expert knowledge. Automatic feature extraction may result in new, more powerful features. In our opinion, it is important to explore the potential ways of ECoG-based movement prediction beyond scalograms.

Another research direction that seems promising is associated with the following hypothesis. We assume that the way spatial pattern of neural activations changes during the movement describes the movement. Distinguishing these spatial-temporal patterns should help in predicting the movement more accurately.

6. Appendix: derivation of NQPFS

To obtain the alternate solution we exploit the fact that the matrix $\underline{\mathbf{A}}$ is binary. Thus an exact low-rank decomposition

$$\underline{\mathbf{A}} = \sum_{r=1}^{R} \mathbf{a}_1^{(r)} \circ \mathbf{a}_2^{(r)} \circ \mathbf{a}_3^{(r)}, \quad \mathbf{a}_1^{(r)} \in \mathbb{R}^{n_1},\ \mathbf{a}_2^{(r)} \in \mathbb{R}^{n_2},\ \mathbf{a}_3^{(r)} \in \mathbb{R}^{n_3} \tag{16}$$

is possible for some R. This allows to rewrite the loss function from (10) as

$$\sum_{r=1}^{R} ||\mathbf{a}_2^{(r)}||_2^2 \cdot ||\mathbf{a}_3^{(r)}||_2^2 \cdot \mathbf{a}_1^{(r)\intercal} \mathbf{Q}_1 \mathbf{a}_1^{(r)} + ||\mathbf{a}_1^{(r)}||_2^2 \cdot ||\mathbf{a}_3^{(r)}||_2^2 \cdot \mathbf{a}_2^{(r)\intercal} \mathbf{Q}_2 \mathbf{a}_2^{(r)} + \tag{17}$$

$$||\mathbf{a}_1^{(r)}||_2^2 \cdot ||\mathbf{a}_2^{(r)}||_2^2 \cdot \mathbf{a}_3^{(r)\intercal} \mathbf{Q}_3 \mathbf{a}_3^{(r)} - \underline{\mathbf{B}} \times_1 \mathbf{a}_1^{(r)} \times_2 \mathbf{a}_2^{(r)} \times_3 \mathbf{a}_3^{(r)}.$$

This problem solves iteratively, via alternate approach. At each step a quadratic programming with problem is solved. Let $\boldsymbol{\alpha}_i = [\mathbf{a}_i^{(1)\intercal}, \ldots, \mathbf{a}_i^{(R)\intercal}]^\intercal \in \mathbb{R}^{nR}$ 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:

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

   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 $\tilde{\mathbf{I}}_1^{(k-1)}$:

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

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

   and

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

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.

[1] S. Amiri, R. Fazel-Rezai, V. Asadpour, A review of hybrid brain-computer interface systems, Advances in Human-Computer Interaction (2013). doi: 10.1155/2013/187024

[2] Y. Bengio, A. Courville, P. Vincent, Representation Learning: A Review and New Perspectives (2012). arXiv:1206.5538

[3] L. Bougrain, N. Liang, Band-specific features improve finger flexion prediction from ECoG, HAL (2009). https://hal.inria.fr/inria-00408675

[4] R. Bro, Multiway calibration. Multilinear PLS, J Chemometr 10(1) (1996) pp. 47–61.

[5] D. T. Bundy, M. Pahwa, N. Szrama, E. C. Leuthardt, Decoding threedimensional reaching movements using electrocorticographic signals in humans, J Neural Eng 13(2) (2016) 026021. doi: 10.1088/1741-2560/13/2/026021

[6] Z. C. Chao, Y. Nagasaka, N. Fujii, Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkeys, Frontiers in Neuroengineering (2010). doi: 10.3389/fneng.2010.00003

[7] Chen C, Shin B, Watanabe H, Nakanishi Y, Kambara H (2013) Prediction of hand trajectory from electrocorticography signals in primary motor cortex. PLoS ONE 8(12):0e83534. doi: 10.1371/journal.pone.0083534

[8] A. Cichocki, Tensor decompositions: A new concept in brain data analysis? Journal of Control, Measurement, and System Integration, special issue: Measurement of Brain Functions and Bio-Signals, 7 (2011) pp. 507–517.

[9] A. Eliseyev, C. Moro, T. Costecalde, N. Torres, S. Gharbi, C. Mestais, A. L. Benabid, T. Aksenova, Iterative N-way partial least squares for a binary self-paced brain–computer interface in freely moving animals, J Neural Eng 88(4) (2011) 046012. doi: 10.1088/1741-2560/8/4/046012

[10] A. Eliseyev, T. Aksenova, Penalized multi-way partial least squares for smooth trajectory decoding from electrocorticographic (ECoG) recording, PLoS ONE 11(5)(2016) e0154878. doi: 10.1371/journal.pone.0154878

[11] A. M. Katrutsa, V. V Strijov, Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria, Expert Syst Appl 76 (2017) pp. 1–11. doi: 10.1016/j.eswa.2017.01.048

[12] J. Kubánek, K. J. Miller, J. G. Ojemann, J. R. Wolpaw, G. Schalk, Decoding flexion of individual fingers using electrocorticographic signals in humans, J Neural Eng 6(6)(2009) 066001. doi: 10.1088/1741-2560/6/6/066001

[13] E. C. Leuthardt, K. J. Miller, G. Schalk, R. P. N. Rao, J. G. Ojemann, Electrocorticography-Based Brain Computer Interface — The Seattle Experience, IEEE T Neur Sys Reh, 14(2) (2006) pp. 194–198.

[14] L. F. Nicolas-Alonso, J. Gomez-Gil, Brain computer interfaces, a review, Sensors 12(2) (2012) pp. 1211–1279. doi: 10.3390/s120201211

[15] K. Shimoda, Y. Nagasaka, Z. C. Chao, N. Fujii, Decoding continuous threedimensional hand trajectories from epidural electrocorticographic signals in Japanese macaques, J Neural Eng, 9 (2012) 036015.

[16] Y. Wang, S. Gao, X. Gao, Common spatial pattern method for channel selection in motor imagery based Brain-computer Interface, IEEE Eng Med Biol 27th Annual Conference (2005). 10.1109/IEMBS.2005.1615701

[17] Q. Zhao, G. Zhou, T. Adali, A. Cichocki. Kernelization of tensor-based models for multiway data analysis. IEEE Signal Proc Mag, 30(4) (2013) pp. 137–148. doi: 10.1109/MSP.2013.2255334

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