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

Anastasia Motrenko

Moscow Institute of Physics and Technology, Russia

Vadim Strijov

Moscow Institute of Physics and Technology, Russia

Abstract

The paper addresses the problem of designing Brain-Computer Interfaces. It investigates feature selection methods in regression, applied to ECoG-based motion decoding. The problem is to predict hand trajectories from the voltage time series of cortical activity. A special characteristic of this problem is the inherently multi-way structure of feature description. The feature description resides in spatial-spectra-temporal domain and includes the voltage time series and their spectral representation. Since electrocorticographic data is highly correlated in temporal, spectral and spatial domains, redundancy of the feature space as well as its dimensionality become a major obstacle for robust solution of the regression problem both in multi-way and flat cases. Feature selection reduces dimensionality and increases model robustness. It plays the crucial role in obtaining adequate predictions.

The main contribution of this paper is the following. We propose a filtering feature selection for multi-way data. The proposed method extends quadratic programming feature selection (QPFS) approach. QPFS selects a subset of features by solving a quadratic problem. It incorporates estimates of similarity between features and their relevance to the regression problem. QPFS offers an effective way to leverage similarities between features and their importance. Our modification allows to apply QPFS to multi-way data. By taking the multi-way structure of features into account, the proposed modification reduces computational costs of optimization problem in QPFS. Exerimental outcomes demonstrate that the proposed modification improves prediction quality of resultant models. The proposed method is model-free and provies interpretable results, which makes it relevant for knowledge extraction and domain analysis.

Keywords: feature selection, brain-computer interface, decoding electrocorticographic data, multi-way data, hand movement prediction

Email addresses: anastasia.motrenko@phystech.edu (Anastasia Motrenko), strijov@ccas.ru (Vadim Strijov)

1. Introduction

We propose a multi-way formulation of a recent approach to filtering feature selection by Katrutsa (Katrutsa & Strijov, 2017), Quadratic Programming Feature Selection (QPFS). A special feature of QPFS is the ability to consider the relationships between features. QPFS is formulated as a quadratic program which minimizes correlation between features while maximizing feature relevance. The paper (Katrutsa & Strijov, 2017) provides extensive comparison of QPFS to LARS, Lasso, Stepwise, Ridge and feature selection with genetic algorithm. The test data displayed various kinds of configurations of features and target vectors. QPFS was shown to outperform the alternatives according to a number of criteria: VIF, BIC and regression quality.

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 for higher dimensionalities. 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.

The main contributions of this paper are:

1. We propose a modification a QPFS, which preserves its attractive properties and also considers multi-way data structure.
2. We apply the proposed method to the problem of hand trajectory reconstruction and show that the quality of reconstruction improves with our method.
3. We compare regression results, obtained with the proposed feature selection method and with alternative regression techniques, according to several criteria and show that the results of the proposed method are at least as accurate.

2. Problem statement

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

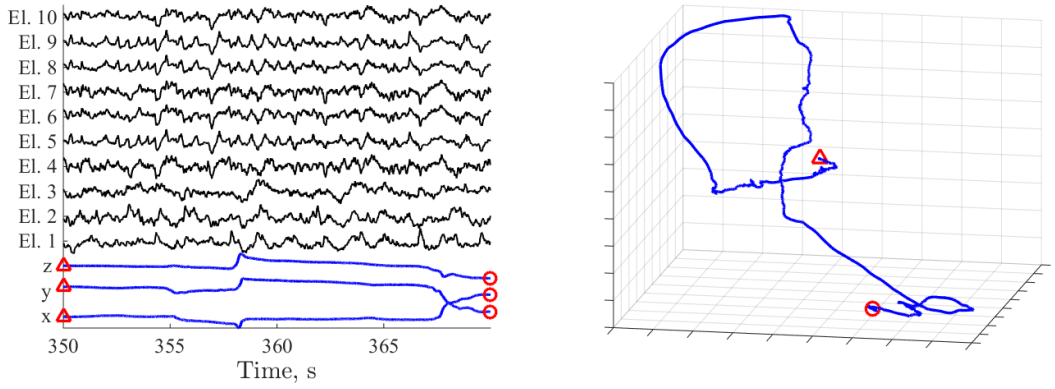
$$\underline{\mathbf{D}} \in \mathbb{R}^{M \times T \times F \times N_{\text{ch}}},\ D_{(m,:,:,:)} = \underline{\mathbf{X}}_m, \quad \mathbf{Y} = [\mathbf{y}_1^{\mathsf{T}}, \ldots, \mathbf{y}_M^{\mathsf{T}}]^{\mathsf{T}}, \tag{1}$$

where each observation corresponds to a certain wrist position $\mathbf{y}_m = \mathbf{y}(t_m)$ and is described by a three-way matrix $\underline{\mathbf{X}}_m \in \mathbb{R}^{T \times F \times N_{\text{ch}}}$ such that 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

[1]The dataset (Shimoda et al., 2012) 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.

Table 1: Multi-way feature selection methods.

Method	Type
(Cao et al., 2014): feature selection based on tensor SVM weights	Wrapper
(Smalter et al., 2009): feature selection based on tensor SVM weights with sparsity induced either with quadratic programming or regularization	Wrapper
(Li & Zhang, 2009): regularized logistic regression	Wrapper
(Zhao et al., 2013): kernelized PLS	Embedded
(Kim et al., 2007): canonical correlation analysis	Embedded / Feature transform
(Eliseyev et al., 2011): iterative multi-way PLS	Embedded / Feature transform
(Eliseyev & Aksenova, 2016): multi-way PLS with additional regularization	Embedded / Feature transform



(a) (b)

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

channel n, $n = 1, \ldots, N_{\text{ch}}$. Extracts from raw ECoG series and target time series are shown in Fig. 1. Fig. 2 illustrates the process of feature extraction. The procedure of feature extraction $\mathbf{s}(t) \to \underline{\mathbf{X}}_m$ will be described in more detail later (see Section 4).

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 $\mathbf{Y}$ with a linear combination of features:

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

where $\text{vec}\big(\underline{\mathbf{X}}_m\big) \in \mathbb{R}^{T \cdot F \cdot N_{\text{ch}}}$ is the result of vectorizing $\underline{\mathbf{X}}_m$, and the weight vector $\hat{\mathbf{w}} \in \mathbb{R}^{T \cdot F \cdot N_{\text{ch}}}$ minimizes 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 correlated in temporal, spatial and spectral domains, linear regression (2) will produce unstable results. To decrease computational cost of the problem (3) and increase robustness of the model we apply feature selection to $\underline{\mathbf{D}}$.

Let $\boldsymbol{\chi}_{ijk} \in \mathbf{R}^M$, $(i, j, k) \in \{1, \ldots, T\} \times \{1, \ldots, F\} \times \{1, \ldots, N_{\text{ch}}\}$ comprise the values of $(i, j, k)-$th feature for all M observations in the data sample. Let $\mathbf{X} \in \mathbb{R}^{M \times T \cdot F \cdot N_{\text{ch}}}$ denote the result of flattening feature matrix $\underline{\mathbf{D}} \in \mathbb{R}^{T \times F \times N_{\text{ch}} \times M}$:

$$\mathbf{X} = [\text{vec}(\underline{\mathbf{X}}_1)^{\intercal}, \ldots, \text{vec}(\underline{\mathbf{X}}_M)^{\intercal}]^{\intercal} = [\ldots, \boldsymbol{\chi}_{ijk}, \ldots]. \tag{4}$$

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. The corresponding two-way feature matrix is

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

Feature selection problem formulates as follows:

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

Here $\mathcal{L}(\hat{\mathbf{Y}}, \mathbf{Y})$ is some loss function and $\mathbf{w}_{\underline{\mathbf{A}}}$ minimizes quadratic loss (3) for $\mathbf{X}_{\underline{\mathbf{A}}}$.

Quality criteria. To evaluate forecasting quality, we used several criteria:

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

$$\text{corr}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{M}\sum_{m=1}^{M} \frac{\text{cov}(\hat{\mathbf{y}}_m, \mathbf{y}_m)}{\sqrt{\text{cov}(\hat{\mathbf{y}}_m, \hat{\mathbf{y}}_m)\text{cov}(\mathbf{y}_m, \mathbf{y}_m)}}. \tag{5}$$

- Scaled MSE

$$\text{sMSE}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{\sum_{m=1}^{M} \|\hat{\mathbf{y}}_m - \mathbf{y}_m\|_2}{\sum_{m=1}^{M} \|\bar{\mathbf{y}} - \mathbf{y}_m\|_2}, \quad \bar{\mathbf{y}} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{y}_m. \tag{6}$$

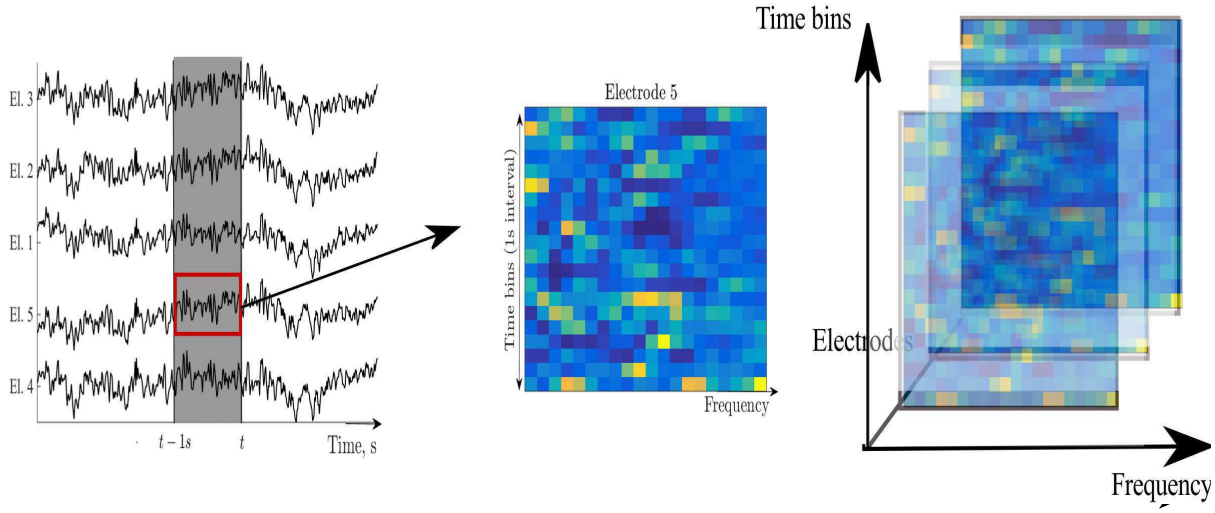


Figure 2: Example of feature construction procedure. For each electrode a one-second long historical interval $[t_m - \Delta t, t_m]$ undergoes wavelet transformation (14) and thus obtains feature description in spectral-temporal domain. Merging spectral-temporal feature matrices for all electrodes, one obtains 3D feature description $\underline{\mathbf{X}}_m$ for the time point t_m.

3. Quadratic Programming Feature Selection

We consider a filtering feature selection approach, proposed in (Katrutsa & Strijov, 2017). Filtering approaches assign individual scores to each variable and select features according to the assigned scores. This does not require model training. Thus, filtering methods are generally more fast than embedded or wrapper methods. However, since filtering methods do not consider relationships between variables, they tend to select correlated features. The advantage of quadratic programming feature selection (QPFS) approach, proposed in (Katrutsa & Strijov, 2017) is that it considers both relevance for prediction and similarity between features. The rest of this Section is outlined as follows. We first explain the original QPFS method, which we refer to as unfolded QPFS. Then we present the proposed multi-way formulation of QPFS. Table 3 summarises notation for multi-way and unfolded feature selection.

3.1. Unfolded QPFS

In case of flat feature description, QPFS problem is formulated as a quadratic program

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

Table 2: Notation for multi-way arithmetics.

$\mathbf{A} \in \mathbb{R}^{m \times n}$	(two-way) matrix
$\mathbf{a} \in \mathbb{R}^{m}$	column vector
$\underline{\mathbf{A}} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$	d-way matrix
$\mathbf{a} \circ \mathbf{b} \in \mathbb{R}^{n_1 \times n_2}$	outer product of $\mathbf{a} \in \mathbb{R}^{n_1}$ and $\mathbf{b} \in \mathbb{R}^{n_2}$: $[\mathbf{a} \circ \mathbf{b}]_{ij} = a_i b_j$
$\underline{\mathbf{A}} \times_1 \mathbf{B} \in \mathbb{R}^{m \times n_2 \times n_3}$	inner product of multi-way matrix $\underline{\mathbf{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ to matrix $\mathbf{B} \in \mathbb{R}^{m \times n_1}$: $[\underline{\mathbf{A}} \times_1 \mathbf{B}]_{ijk} = \sum_{i'} a_{i'jk} b_{ii'}$
$\underline{\mathbf{A}} * \underline{\mathbf{B}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$	element-wise product of multi-way $\underline{\mathbf{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and $\underline{\mathbf{B}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$: $[\underline{\mathbf{A}} * \underline{\mathbf{B}}]_{ijk} = a_{ijk} b_{ikj}$

where $(i, j)-$th entry q_{ij} of matrix $\mathbf{Q} \in \mathbb{R}^{N \times N}$ quantifies *similarity* between i-th and j-th features. We use pairwise correlation coefficient to measure similarity between features:

$$q_{ij} = |\text{corr}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j). \tag{8}$$

Similarly, we measure *relevance* b_i of the $i-$th feature as its correlation with the target $\mathbf{Y}$:

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

where $\boldsymbol{y}_n$ are columns of the target matrix $\mathbf{Y}$. The paper (Katrutsa & Strijov, 2017) considers other ways to define $\mathbf{Q}$ and $\mathbf{b}$, such as mutual information and normalized feature significance as similarity and relevance measures.

The problem (7) balances between increasing diversity of the selected features and maximizing their predictive power. This is done through optimization by a binary vector $\mathbf{a} \in \mathbb{R}^N$, which defines the active set of predictors:

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

3.2. Multi-way QPFS

One way to address feature selection in multi-way case is to flatten the features $\underline{\mathbf{D}}$ by vectorizing (4) each three-way[2] matrix $\underline{\mathbf{X}}_m$ and then proceed with the original QPFS (7). The weak point of such approach is computing similarity matrix $\mathbf{Q} \in \mathbb{R}^{n_1 n_2 n_2 \times n_1 n_2 n_3}$. Since $\mathbf{X}$ contains multiple correlated features, $\mathbf{Q}$ becomes closer to singular as the number of features grows, even though it is positive-definite by definition (8). The construction of the optimization problem becomes the most difficult part of QPFS. To overcome this problem,

[2]We formulate multi-way QPFS for the case of four-way data $\underline{\mathbf{D}} \in \mathbb{R}^{M \times n_1 \times n_2 \times n_3}$ $d = 3$ (three-way features), but all the derivations generalize to any number of modes $\underline{\mathbf{X}}_m \in \mathbb{R}^{n_1 \times \cdots \times n_d}, d \geq 2$.

Table 3: Correspondence between flat (two-way) and multi-way feature descriptions.

	Flat case	Multiindex case
Design matrix	$\mathbf{X} \in \mathbb{R}^{M \times n}$	$\underline{\mathbf{D}} \in \mathbb{R}^{M \times n_1 \times n_2 \times n_3}$
Vector of feature values	$\boldsymbol{\chi}_i \in \mathbb{R}^n$	$\underline{\mathbf{X}}_{(:,i_1,i_2,i_3)} \in \mathbb{R}^{n_1 n_2 n_3}$
Indicator variables	$\mathbf{a} \in \mathbb{R}^n$	$\underline{\mathbf{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$
Similarity	$\mathbf{Q} \in \mathbb{R}^{n \times n}$	$\mathbf{Q}_1 \in \mathbb{R}^{n_1 \times n_1}$, $\mathbf{Q}_2 \in \mathbb{R}^{n_2 \times n_2}$, $\mathbf{Q}_3 \in \mathbb{R}^{n_3 \times n_3}$
Relevance	$\mathbf{b} \in \mathbb{R}^n$	$\underline{\mathbf{B}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$

we incorporate the multi-way structure of ECoG features $\underline{\mathbf{X}}_m$ into feature selection problem 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 $\underline{\mathbf{X}}_m \in \mathbb{R}^{n_1 \times n_2 \times n_3}$.

To proceed further, we need to introduce notation for multi-way arithmetics.

- Let $\mathbf{a} \circ \mathbf{b}$ denote the outer product of two vectors $\mathbf{a} \in \mathbb{R}^{n_1}, \mathbf{b} \in \mathbb{R}^{n_2}$

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

- Let $\underline{\mathbf{A}} \times_d \mathbf{B}$ 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 $\mathbf{B} \in \mathbb{R}^{m \times n_1}$

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

- Let $\underline{\mathbf{A}} * \underline{\mathbf{B}}$ 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 matrix $\underline{\mathbf{B}}$ are known. The problem (7) reformulates as follows:

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

where 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 $\underline{\mathbf{A}}$.

$$\underline{\mathbf{A}} = \sum_{r=1}^{R} \mathbf{a}_1^{(r)} \circ \mathbf{a}_2^{(r)} \circ \mathbf{a}_3^{(r)}, \quad \mathbf{a}_1^{(r)} \in \mathbb{R}^{n_1}, \ \mathbf{a}_2^{(r)} \in \mathbb{R}^{n_2}, \ \mathbf{a}_3^{(r)} \in \mathbb{R}^{n_3}, \tag{11}$$

This decomposition allows to solve the problem (10) via alternate approach, so that at each step a quadratic program is solved. The derivation and exact formulation of the proposed multi-way QPFS algorithm can be found in AppendixA.

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

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

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

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

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

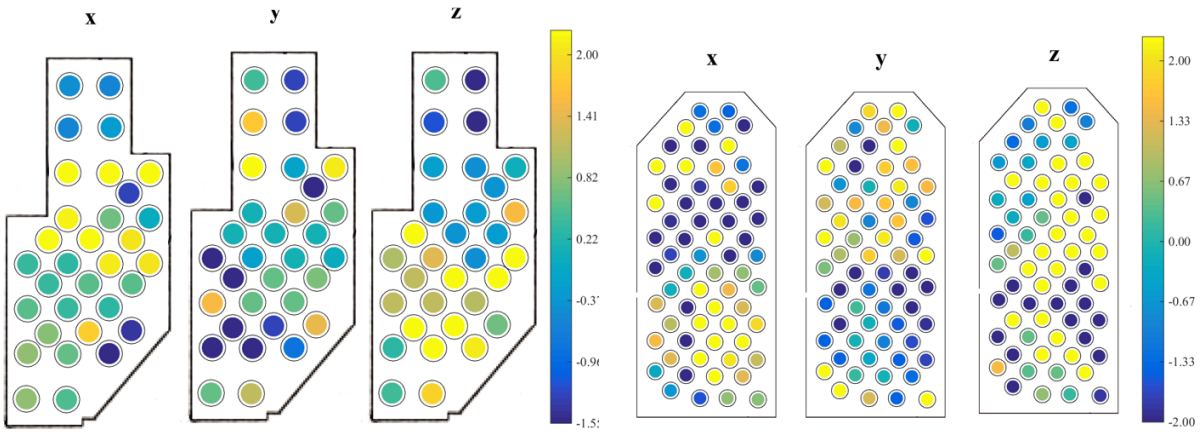
Linear relaxation of (10). The problem (10) is the integer optimization problem, which is not convex. To allow for more efficient solution, we have to relax non-convex problem constraint $\underline{\mathbf{A}} \in \{0, 1\}^{n_1 \times n_2 \times n_3}$ into $\hat{\underline{\mathbf{A}}} \in [0, 1]^{n_1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

$$\underline{\mathbf{A}}(\epsilon) = [a_{ijk}], \quad a_{ijk} = \begin{cases} 1 \text{ if } \hat{a}_{ijk} \geq \epsilon, \\ \text{otherwise} \end{cases} \tag{12}$$

Thresholded solution $\underline{\mathbf{A}}(\epsilon)$ defines an active set 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

$$\boldsymbol{\chi}_{ijk} \preceq \boldsymbol{\chi}_{i'j'k'} \Leftrightarrow \hat{a}_{ijk} \leq \hat{a}_{i'j'k'}. \tag{13}$$

Complexity. Solving a convex quadratic program (7) takes polynomial time in $n_1 \cdot n_2 \cdot n_3$. To formulate the problem, one needs to compute $(n_1 \cdot n_2 \cdot n_3)^2$ entries q_{ij} of Q, which might become prohibitive. In the proposed algorithm, one makes several iteration, at each iteration solving three convex quadratic programs (A.3). Each problem requires computation of n_d^2 values, which is way less than $(n_1 \cdot n_2 \cdot n_3)^2$. Solving each problem takes polynomial time in $n_d \cdot R$. Here R is bounded by the number of desired sparse feature set: $n_d \cdot R \leq n_d^2$. Even multiplied by a number of iterations, $\mathcal{O}\left((n_1^2)^{p_1}\right) + \mathcal{O}\left((n_2^2)^{p_2}\right) + \mathcal{O}\left((n_3^2)^{p_3}\right)$ is still asymptotically less than $\mathcal{O}\left((n_1 \cdot n_2 \cdot n_3)^p\right)$. In practice, we found that using only one iteration was usually enough.



(a) A (b) K1

Figure 3: Optimal latency values for monkeys A (32 electrodes) and K1 (64 electrodes). Each circle corresponds to one electrode. Positions of the circles roughly describe electrode placements.

4. Experiments

Feature extraction for ECoG data. We used NeuroTycho foodtracking dataset (Shimoda et al., 2012) for evaluation. ECoG signals and wrist positions were measured simultaneously. The dataset includes observations for two monkeys, A ($N_{\text{ch}} = 32$) and K ($N_{\text{ch}} = 64$) across several dates. The voltage time series $\mathbf{s}(t)$ were sampled at 1000Hz, wrist positions were sampled at 120Hz.

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 (Kubánek et al., 2009; Bougrain & Liang, 2009; Bundy et al., 2016). The feature description includes frequency- and time-domain features. Frequency-domain features are obtained with spectral transform. These features represent time-dependent contributions of a range of frequencies into the signal.We use wavelet transform to obtain spectro-temporal features. A comparison (Lotric et al., 2000)of spectral analysis methods showed that wavelet transforms provide better frequency resolution than Short Time Fourier transform or autoregressive analysis. We used Morlet wavelet as mother wavelet, since it is commonly used in BCI data analysis (Chao et al., 2010; Eliseyev & Aksenova, 2016; **?**). The time-domain features, referred to as local motor potentials (Kubánek et al., 2009), are essentially low-passed ECoG time series $\mathbf{s}(t)$. Both temporal and spectral features are time delayed.

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

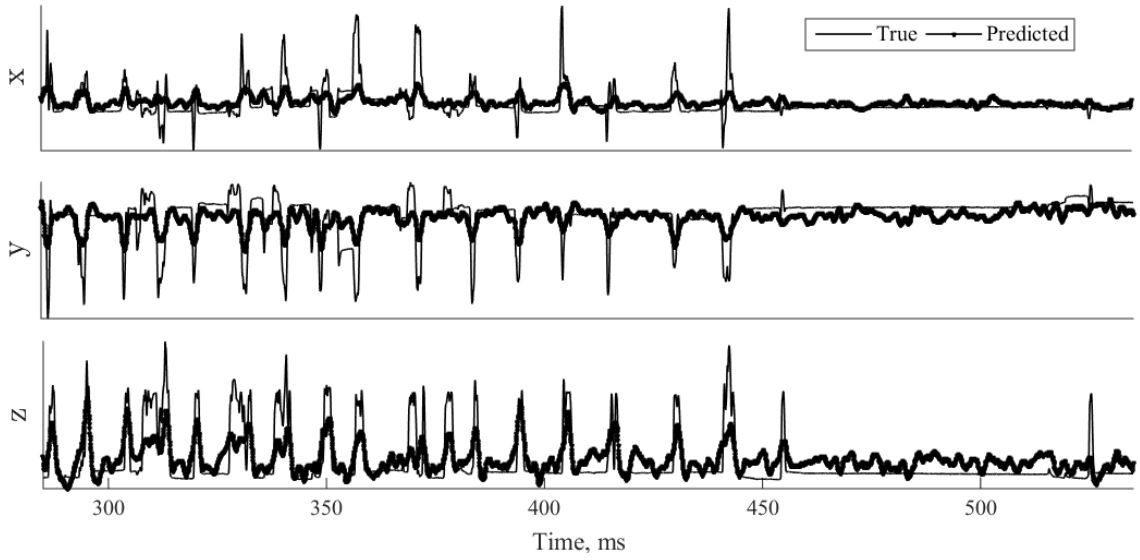


Figure 4: Segment of the forecasted time series. Linear regression, 50 best features according to multi-way QPFS.

Optimal latency τ^* values depend on the electrode position and the spatial pattern of this dependency varies between x, y, z dimensions of target time series.

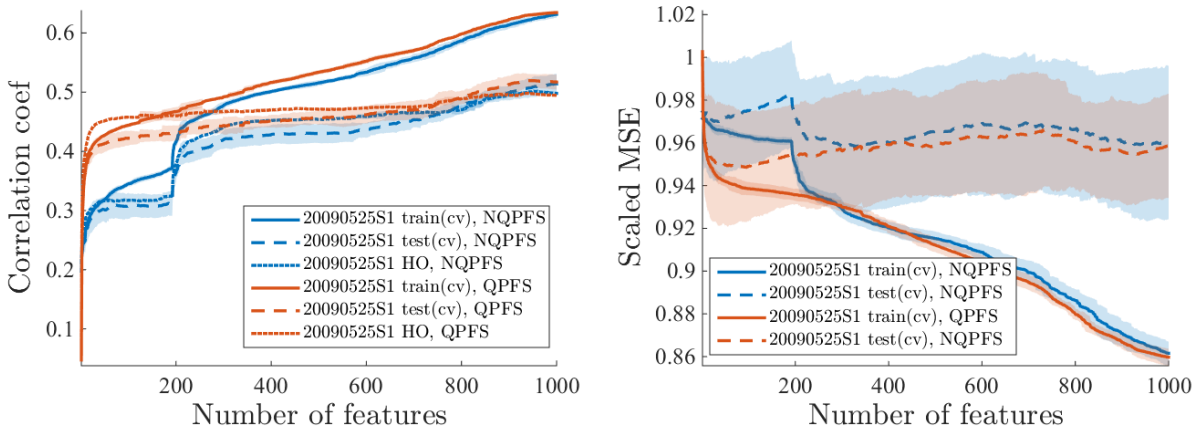
Frequency-domain features. The feature matrix $\underline{\mathbf{X}}_m$ comprises spatial, temporal and spectral information about the time series $\mathbf{s}(t)$ across the time period $[t_m - \Delta, t_m]$. Fig. 2 illustrates the process of feature extraction. The spatial component is represented 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 transform 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 $t_1 \leq t_i \leq t_M$ and scale f_j, $j = 1, \ldots, F$:

$$W_{ijn} = \frac{1}{\sqrt{|f_j|}} \sum_{t \leq t_M} \psi\left(\frac{t - t_i}{f_j}\right) s_n(t). \tag{14}$$

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:

$$\underline{\mathbf{X}}_m \in \mathbb{R}^{F \times N_{\text{ch}}}, \quad \underline{\mathbf{X}}_{mjn} = \begin{cases} s_n(t_m + \tau), j = 1, \\ W_{mjn} \text{ for } j = 2, \ldots, F+1, \end{cases} \quad n = 1, \ldots, N_{\text{ch}}. \tag{15}$$



(a) (b)

Figure 5: Forecasting quality by model complexity. Features are added by one in order (13) defined by QPFS. The quality is measured as (a) the correlation coefficient and (b) scaled MSE between the predicted and the true wrist trajectory.

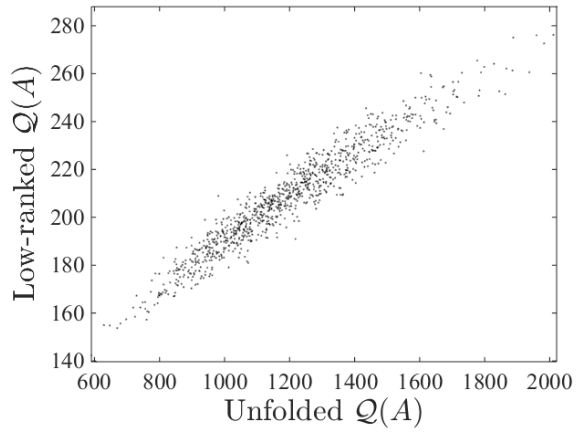


Figure 6: Values of loss function of unfolded QPFS against values loss function for multi-way QPFS.

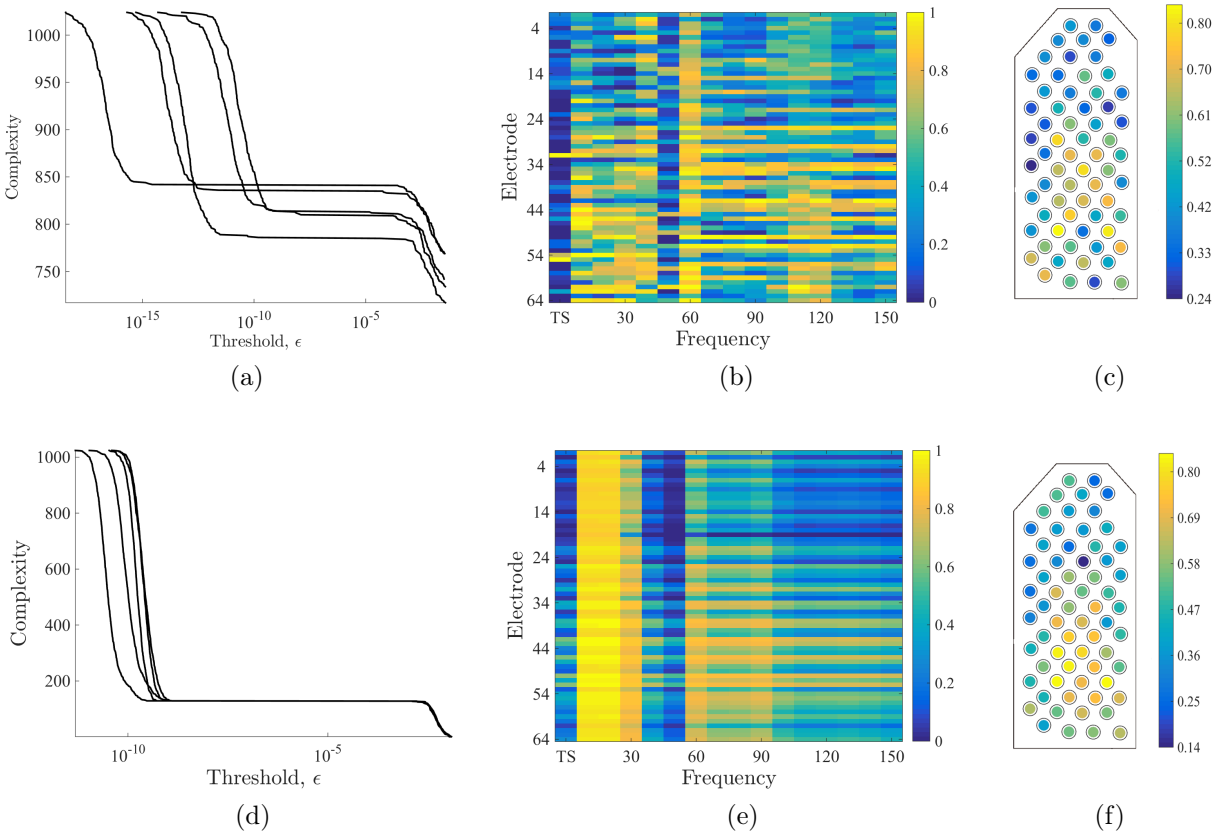


Figure 7: (a) Complexity by the threshold value ϵ. Each line corresponds to a cross-validation split. (b) Evaluation of electrode-frequency pairs importance. Importance is measured as feature rank (13), averaged over cross-validation splits. (c) Electrode ranks, averaged over frequencies.

The time series were downsampled the data by the factor of 10. To create the dataset 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 $|t_i \in \Delta_m| \geq T$, 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, $i = 1, \ldots, T$. 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 δt_i:

$$X_{mijn} = \frac{1}{|\delta t_i|} \sum_{i' \colon t_{i'} \in \delta t_i} W_{i'jn}. \tag{16}$$

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. In this section we compare performance of original QPFS and multi-way QPFS[3]. Fig. 7 summarizes results of multi-way QPFS, applied to the 2D feature set (15). To evaluate performance of the QPFS algorithm, we split the part of the dataset, 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 ranking of features (13). We say that the feature is ranked $n-$th, if it is worse than exactly $n - 1$ features. Since higher ranked features are more likely to be included into the model, we measured feature utility as its rank, averaged over cross-validation splits.

To obtain an active feature set $\mathbf{X}_{\underline{\mathbf{A}}(\epsilon)}$ we 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 (5) or (6) of linear model (2), with parameters $\mathbf{w}_{\underline{\mathbf{A}}(\epsilon)}$ estimated at the training set.

Fig. 4, 5, and 7 exemplify QPFS results for 2D feature set. Fig. 4 demonstrates an example of a forecast, obtained for wrist trajectory with 50 features. The features were selected by multi-way QPFS from 2D feature set. As can be seen, the reconstructed trajectory follows the peaks in the original trajectory. However, it is too jerky in the "still" regions (450 ms and further), 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

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

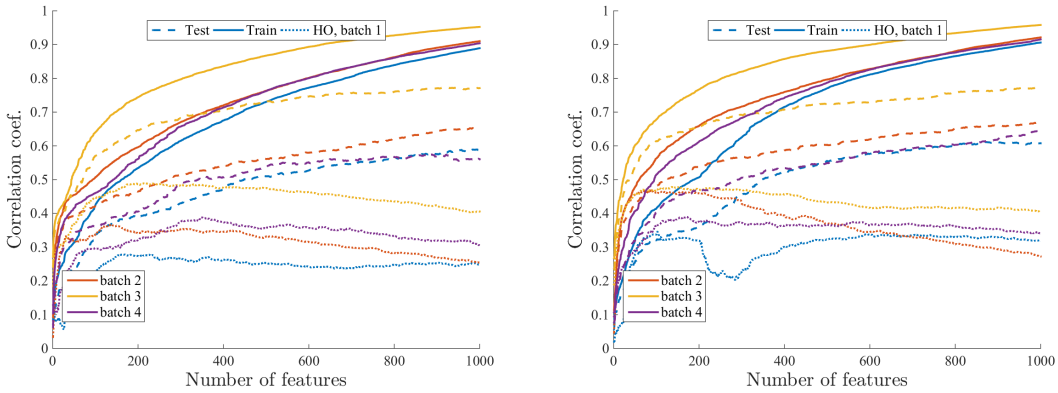


Figure 8: Forecasting quality measures as correlation coefficient between the original wrist trajectory and the reconstructed trajectory. (a) Unfolded QPFS. (b) Multi-way QPFS.

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 compares values of loss function of unfolded QPFS (7) and the corresponding values of multi-way QPFS (10). Each point on the figure corresponds to a subset of features. Feature sets were selected at random. Unfolded and multi-way loss functions demonstrate strong positive correlation, which implies that minimizer of (10) should yield value of (7) close to its minimum.

Fig. 7(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. 7(b) color-codes utility of each feature (electrode-frequency pair). The first column of the color-coded matrix corresponds to decimated ECoG time series. Fig. 7(c) shows color-coded utility 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. 7 demonstrates that multi-way and unfolded QPFS tend to select different feature sets. Multi-way QPFS selects less diverse features, since there it makes no explicit pairwise comparison.

In case of 3D feature set, the number of features becomes prohibitive for unfolded QPFS. For these reason, all comparisons made in this paper, are based on 2D feature set. We were able to use multi-way QPFS in case of 3D feature selection, but we had to additionally split the data into batches. Fig. 8 shows how quality curves change their shape from batch to batch.

Comparison of QPFS and PLS. As a reference embedded algorithm we selected PLS regression, since it is reported successful in ECoG-based motion reconstruction. We compared the proposed algorithm to unfolded PLS and NPLS, the way it was formulated in (Eliseyev et al., 2011). Table 4 compares unfolded QPFS, multi-way QPFS (labeled NQPFS by analogy with NPLS), PLS and NPLS in terms of correlation coefficient (5). Each algorithm was allowed to select $N = 10, 25, 200$ or 500 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 dataset (5 datasets for monkey A and 3 datasets for monkey K) and each N the best result is given in bold.

Besides NeuroTycho dataset we considered gestures datasets from Stanford Digital repository (?). The datasets contain electrocorticographic data and finger flexions for finger movement prediction. Three datasets were considered:

- 'base': basic baseline fixation task,
- 'fingerflex': the subject move a certain finger upon a cue,
- 'freeform': self-paced movements in response.

Recordings for each subject lasted for about 2 minutes. The results are listed in the tables 5, 6, 7. Due to relatively small sample size we used smaller number of selected features for estimation.

In addition to previously mentioned correlation coefficient (5) and scaled MSE (6) we used dynamic time warping (DTW) 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 aligned before comparison and represents the cost of best alignment. MADE measures smoothness of the reconstructed trajectory as

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

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

Finally, we assess how time-consuming are the compared algorithms (see Fig. 10). Since PLS and NPLS simultaneously reduce dimensionality and train the model, we included the time needed to train linear model (3) into the estimates for QPFS and NQPFS. We considered $\underline{\mathbf{D}} \in \mathbb{R}^{5000 \times n \times n}$ for $n \in \{10, 25, 50\}$. Fig. 10(a) shows time estimates against the number of selected features with no model training for QPFS and NQPFS. Since QPFS and NQPFS rank all features simultaneously, there is no dependance on the number of

Table 4: Correlation coefficient between predicted and true wrist trajectory.

Monkey, date	Algorithm	$N = 10$	$N = 25$	$N = 200$	$N = 500$
A, 20090116	QPFS	$\mathbf{0.333 \pm 0.007}$	0.34 ± 0.008	0.414 ± 0.041	0.505 ± 0.015
	NQPFS	0.271 ± 0.028	0.307 ± 0.009	0.459 ± 0.014	$\mathbf{0.511 \pm 0.020}$
	PLS	0.306 ± 0.007	$\mathbf{0.376 \pm 0.012}$	$\mathbf{0.476 \pm 0.017}$	0.51 ± 0.010
	NPLS	0.0367 ± 0.006	0.125 ± 0.016	0.329 ± 0.021	0.358 ± 0.054
K, 20090525	QPFS	$\mathbf{0.252 \pm 0.079}$	0.288 ± 0.096	0.392 ± 0.144	$\mathbf{0.467 \pm 0.132}$
	NQPFS	0.185 ± 0.040	0.225 ± 0.063	0.374 ± 0.119	0.416 ± 0.112
	PLS	0.24 ± 0.042	$\mathbf{0.288 \pm 0.095}$	$\mathbf{0.429 \pm 0.128}$	0.409 ± 0.122
	NPLS	0.0359 ± 0.026	0.115 ± 0.007	0.278 ± 0.092	0.336 ± 0.072
K, 20090527	QPFS	$\mathbf{0.346 \pm 0.020}$	$\mathbf{0.391 \pm 0.007}$	0.403 ± 0.007	$\mathbf{0.441 \pm 0.006}$
	NQPFS	0.185 ± 0.012	0.191 ± 0.007	0.384 ± 0.006	0.413 ± 0.009
	PLS	0.297 ± 0.010	0.347 ± 0.013	$\mathbf{0.428 \pm 0.018}$	0.417 ± 0.028
	NPLS	0.118 ± 0.019	0.23 ± 0.016	0.354 ± 0.025	0.371 ± 0.018
K, 20090602	QPFS	$\mathbf{0.357 \pm 0.017}$	$\mathbf{0.382 \pm 0.014}$	0.424 ± 0.014	0.468 ± 0.010
	NQPFS	0.29 ± 0.007	0.306 ± 0.003	0.381 ± 0.006	$\mathbf{0.474 \pm 0.017}$
	PLS	0.341 ± 0.012	0.369 ± 0.010	$\mathbf{0.442 \pm 0.008}$	0.416 ± 0.015
	NPLS	0.111 ± 0.009	0.167 ± 0.007	0.306 ± 0.012	0.344 ± 0.026
A, 20081127	QPFS	0.281 ± 0.014	0.323 ± 0.017	0.397 ± 0.028	0.438 ± 0.066
	NQPFS	0.205 ± 0.008	0.227 ± 0.012	$\mathbf{0.436 \pm 0.012}$	$\mathbf{0.451 \pm 0.016}$
	PLS	$\mathbf{0.293 \pm 0.016}$	$\mathbf{0.353 \pm 0.009}$	0.429 ± 0.020	0.434 ± 0.064
	NPLS	0.0402 ± 0.012	0.0698 ± 0.002	0.2 ± 0.011	0.224 ± 0.025
A, 20081224	QPFS	0.262 ± 0.012	0.291 ± 0.035	0.332 ± 0.089	0.4 ± 0.079
	NQPFS	0.196 ± 0.005	0.216 ± 0.003	0.298 ± 0.075	$\mathbf{0.407 \pm 0.086}$
	PLS	$\mathbf{0.276 \pm 0.006}$	$\mathbf{0.331 \pm 0.010}$	$\mathbf{0.349 \pm 0.086}$	0.401 ± 0.090
	NPLS	0.0894 ± 0.010	0.129 ± 0.008	0.169 ± 0.065	0.206 ± 0.082
A, 20090121	QPFS	0.222 ± 0.006	0.242 ± 0.011	0.344 ± 0.027	0.368 ± 0.026
	NQPFS	0.18 ± 0.011	0.224 ± 0.011	$\mathbf{0.362 \pm 0.018}$	$\mathbf{0.371 \pm 0.036}$
	PLS	$\mathbf{0.248 \pm 0.012}$	$\mathbf{0.297 \pm 0.016}$	0.341 ± 0.022	0.369 ± 0.029
	NPLS	0.0262 ± 0.011	0.0641 ± 0.008	0.151 ± 0.014	0.172 ± 0.021
A, 20090611	QPFS	$\mathbf{0.3 \pm 0.017}$	0.321 ± 0.022	0.345 ± 0.011	$\mathbf{0.357 \pm 0.011}$
	NQPFS	0.251 ± 0.005	0.256 ± 0.005	0.326 ± 0.004	0.355 ± 0.005
	PLS	0.282 ± 0.010	$\mathbf{0.321 \pm 0.009}$	$\mathbf{0.359 \pm 0.008}$	0.357 ± 0.011
	NPLS	0.108 ± 0.006	0.184 ± 0.007	0.27 ± 0.010	0.27 ± 0.008
Average	QPFS	$\mathbf{0.294 \pm 0.022}$	0.322 ± 0.026	0.381 ± 0.045	$\mathbf{0.43 \pm 0.043}$
	NQPFS	0.22 ± 0.015	0.244 ± 0.014	0.377 ± 0.032	0.425 ± 0.038
	PLS	0.285 ± 0.014	$\mathbf{0.335 \pm 0.022}$	$\mathbf{0.407 \pm 0.038}$	0.414 ± 0.046
	NPLS	0.0705 ± 0.012	0.135 ± 0.009	0.257 ± 0.031	0.285 ± 0.038

Table 5: Correlation coefficient for gestures 'base' dataset.

Subject	Algorithm	$N = 10$	$N = 25$	$N = 50$	$N = 100$
wm_base	QPFS	0.141 ± 0.058	0.0979 ± 0.067	0.0863 ± 0.064	0.0669 ± 0.075
	NQPFS	$\mathbf{0.23 \pm 0.080}$	$\mathbf{0.227 \pm 0.087}$	0.204 ± 0.089	0.172 ± 0.084
	PLS	0.109 ± 0.068	0.0911 ± 0.068	0.107 ± 0.061	0.0702 ± 0.068
	NPLS	0.218 ± 0.103	0.212 ± 0.101	$\mathbf{0.226 \pm 0.113}$	$\mathbf{0.231 \pm 0.112}$
de_base	QPFS	$\mathbf{0.336 \pm 0.332}$	$\mathbf{0.325 \pm 0.306}$	$\mathbf{0.343 \pm 0.323}$	$\mathbf{0.363 \pm 0.310}$
	NQPFS	0.293 ± 0.310	0.232 ± 0.299	0.212 ± 0.257	0.248 ± 0.342
	PLS	0.118 ± 0.031	0.0453 ± 0.021	0.0383 ± 0.018	0.0978 ± 0.054
	NPLS	0.102 ± 0.032	0.104 ± 0.047	0.0643 ± 0.042	0.0961 ± 0.120
bp_base	QPFS	0.354 ± 0.317	$\mathbf{0.374 \pm 0.310}$	$\mathbf{0.35 \pm 0.319}$	$\mathbf{0.331 \pm 0.286}$
	NQPFS	$\mathbf{0.369 \pm 0.311}$	0.345 ± 0.335	0.233 ± 0.372	0.218 ± 0.377
	PLS	0.128 ± 0.051	0.187 ± 0.043	0.258 ± 0.050	0.32 ± 0.031
	NPLS	0.0422 ± 0.031	0.0983 ± 0.098	0.097 ± 0.118	0.101 ± 0.082
ca_base	QPFS	0.0235 ± 0.017	0.0282 ± 0.022	0.0486 ± 0.046	0.175 ± 0.088
	NQPFS	$\mathbf{0.186 \pm 0.244}$	$\mathbf{0.129 \pm 0.111}$	$\mathbf{0.264 \pm 0.351}$	$\mathbf{0.205 \pm 0.224}$
	PLS	0.0526 ± 0.032	0.0627 ± 0.022	0.0655 ± 0.070	0.0582 ± 0.026
	NPLS	0.0836 ± 0.031	0.116 ± 0.036	0.135 ± 0.028	0.107 ± 0.054
cc_base	QPFS	0.259 ± 0.269	0.232 ± 0.211	0.189 ± 0.217	0.128 ± 0.234
	NQPFS	$\mathbf{0.267 \pm 0.183}$	$\mathbf{0.462 \pm 0.138}$	$\mathbf{0.423 \pm 0.129}$	$\mathbf{0.39 \pm 0.102}$
	PLS	0.0893 ± 0.058	0.057 ± 0.022	0.0482 ± 0.034	0.122 ± 0.037
	NPLS	0.25 ± 0.111	0.315 ± 0.075	0.328 ± 0.097	0.312 ± 0.132

Table 6: Correlation coefficient for gestures 'fingers'.

Subject	Algorithm	$N = 10$	$N = 25$	$N = 50$	$N = 100$
wm_fingerflex	QPFS	0.407 ± 0.029	0.392 ± 0.051	0.339 ± 0.154	0.36 ± 0.092
	NQPFS	$\mathbf{0.424 \pm 0.009}$	0.388 ± 0.068	0.387 ± 0.068	0.404 ± 0.038
	PLS	0.415 ± 0.019	$\mathbf{0.488 \pm 0.019}$	$\mathbf{0.571 \pm 0.016}$	$\mathbf{0.582 \pm 0.030}$
	NPLS	0.395 ± 0.021	0.394 ± 0.017	0.401 ± 0.047	0.348 ± 0.111
bp_fingerflex	QPFS	0.235 ± 0.034	0.239 ± 0.027	0.242 ± 0.026	0.238 ± 0.021
	NQPFS	0.0287 ± 0.020	0.0405 ± 0.019	0.0587 ± 0.015	0.0702 ± 0.018
	PLS	$\mathbf{0.237 \pm 0.007}$	$\mathbf{0.282 \pm 0.006}$	$\mathbf{0.318 \pm 0.014}$	$\mathbf{0.366 \pm 0.013}$
	NPLS	0.0632 ± 0.010	0.0329 ± 0.011	0.0452 ± 0.012	0.0811 ± 0.030
ca_fingerflex	QPFS	0.0811 ± 0.031	0.0457 ± 0.031	0.0883 ± 0.030	0.103 ± 0.045
	NQPFS	0.0824 ± 0.020	0.08 ± 0.038	0.0796 ± 0.041	0.11 ± 0.017
	PLS	$\mathbf{0.263 \pm 0.012}$	$\mathbf{0.284 \pm 0.020}$	$\mathbf{0.284 \pm 0.017}$	$\mathbf{0.274 \pm 0.014}$
	NPLS	0.0963 ± 0.021	0.0128 ± 0.007	0.0225 ± 0.013	0.0202 ± 0.015

Table 7: Correlation coefficient for gestures 'freeform' dataset.

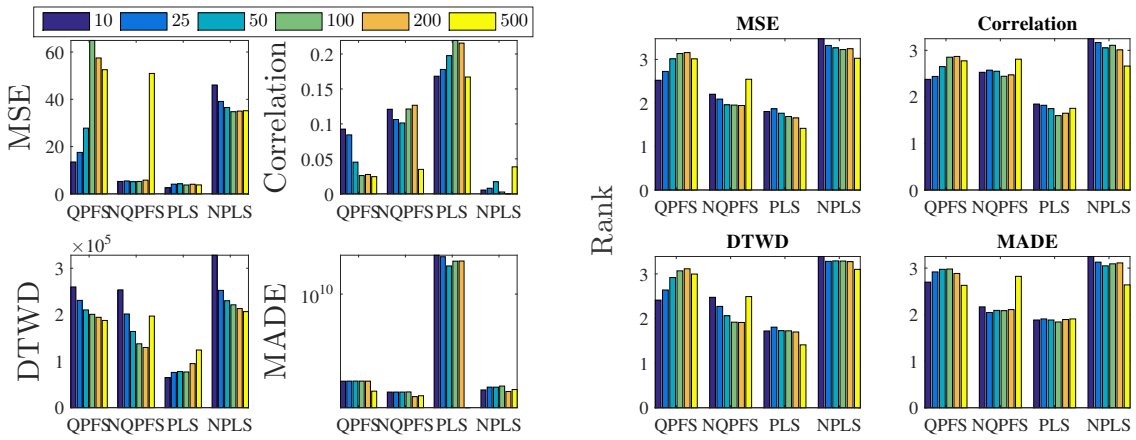
Subject	Algorithm	$N = 10$	$N = 25$	$N = 50$	$N = 100$
wm_freeform	QPFS	0.0626 ± 0.032	0.108 ± 0.041	0.102 ± 0.032	0.121 ± 0.059
	NQPFS	0.345 ± 0.032	0.302 ± 0.060	0.277 ± 0.032	0.314 ± 0.042
	PLS	$\mathbf{0.392 \pm 0.196}$	$\mathbf{0.333 \pm 0.214}$	0.343 ± 0.189	0.406 ± 0.182
	NPLS	0.323 ± 0.152	0.328 ± 0.199	$\mathbf{0.409 \pm 0.192}$	$\mathbf{0.437 \pm 0.210}$
de_freeform	QPFS	$\mathbf{0.32 \pm 0.149}$	$\mathbf{0.339 \pm 0.163}$	$\mathbf{0.355 \pm 0.159}$	0.366 ± 0.169
	NQPFS	0.31 ± 0.059	0.286 ± 0.059	0.228 ± 0.118	0.204 ± 0.117
	PLS	0.297 ± 0.040	0.333 ± 0.009	0.338 ± 0.025	$\mathbf{0.4 \pm 0.054}$
	NPLS	0.243 ± 0.047	0.277 ± 0.034	0.308 ± 0.054	0.357 ± 0.059
bp_freeform	QPFS	0.179 ± 0.086	0.194 ± 0.086	0.127 ± 0.072	0.0352 ± 0.032
	NQPFS	$\mathbf{0.316 \pm 0.032}$	$\mathbf{0.368 \pm 0.045}$	0.101 ± 0.108	0.061 ± 0.027
	PLS	0.238 ± 0.044	0.274 ± 0.038	$\mathbf{0.284 \pm 0.051}$	$\mathbf{0.269 \pm 0.045}$
	NPLS	0.0302 ± 0.022	0.0453 ± 0.020	0.0806 ± 0.031	0.0961 ± 0.028
ca_freeform	QPFS	0.13 ± 0.072	0.101 ± 0.029	0.108 ± 0.045	$\mathbf{0.211 \pm 0.175}$
	NQPFS	$\mathbf{0.252 \pm 0.157}$	0.226 ± 0.172	$\mathbf{0.185 \pm 0.189}$	0.195 ± 0.183
	PLS	0.153 ± 0.051	$\mathbf{0.233 \pm 0.052}$	0.183 ± 0.030	0.174 ± 0.055
	NPLS	0.0724 ± 0.033	0.0686 ± 0.025	0.0843 ± 0.026	0.106 ± 0.130
cc_freeform	QPFS	0.182 ± 0.042	0.194 ± 0.021	0.236 ± 0.058	0.246 ± 0.016
	NQPFS	$\mathbf{0.28 \pm 0.027}$	$\mathbf{0.304 \pm 0.038}$	$\mathbf{0.316 \pm 0.042}$	$\mathbf{0.264 \pm 0.111}$
	PLS	0.155 ± 0.025	0.153 ± 0.013	0.202 ± 0.036	0.24 ± 0.047
	NPLS	0.0415 ± 0.027	0.0403 ± 0.031	0.0478 ± 0.019	0.0503 ± 0.034

selected features. The most consuming part of QPFS computing $n^2 \times n^2$ similarity matrix and verifying it is adequate for further computations. Since NQPFS operates only with two $n \times n$ matrices it performs much better in terms of timing. Fig. 10(b) shows time taken to select features and train the model. We see that QPFS and NQPFS still depend less on the number of selected features than both versions of PLS. Also, is is seen that training linear model with NQPFS is as efficient as training PLS regression model.

5. Conclusion

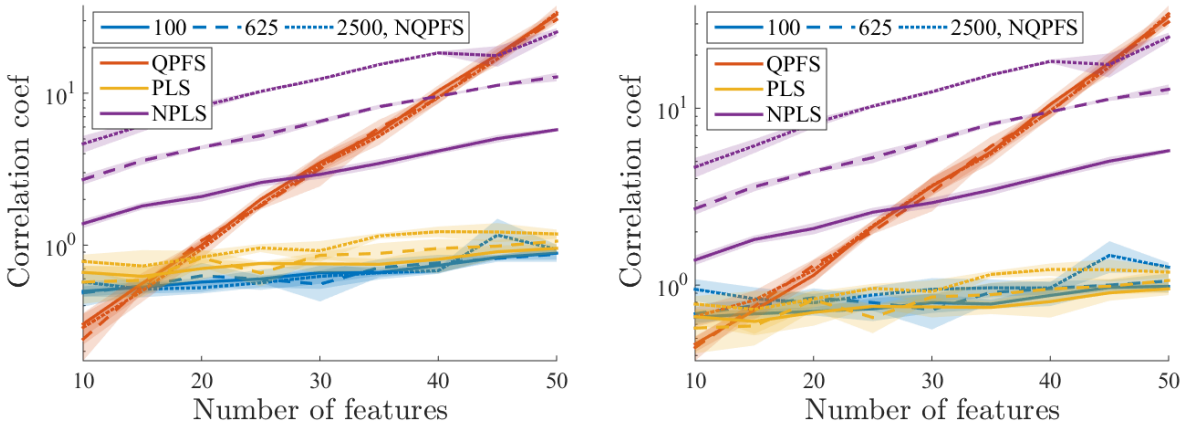
Decoding cortical activity of human brain is the central problem of neuroengineering. We address the problem of 3D movement reconstruction from cortical activity. In order to build a model that is both adequate and computationally simple, we propose a multi-way formulation of the quadratic programming feature selection approach. QPFS is a flexible and efficient approach, which allows to select most relevant features from a highly 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. The quality of simple linear regression based on QPFS-selected features is comparable with the quality PLS regression. We observed that multi-way QPFS produced similar results to those of QPFS in terms of regression quality and did it much more efficiently due to exploitation of multi-way structure of the data.



(a) (b)

Figure 9: (a) Average values of all quality criteria for the compared algorithms. (b) Average rankings of the compared algorithms (the less – the better).



(a) (b)

Figure 10: (a) Algorithm timings, QPFS and NQPFS measured without model training. (b) Algorithm timings with model training.

The proposed method can be used to construct predictive models in Brain-Computer Interface design. We believe that our optimization-based feature selection algorithm will be able to provide efficient solution for feature selection problem in other multi-way settings.

The weaknesses of the proposed method are those of filtering methods. In terms of regression quality wrapper and embedded methods might provide more accurate results, since these methods incorporate feature selection into the fitting process. Also, compared to the original QPFS, the proposed approach tends to select less diverse feature sets, since no pair of features is compared explicitly. To increase diversity one has to introduce additional constant factor to the problem (7) to better leverage between the summands.

The strengths of the proposed approach are the following:

- As a filtering method, the proposed approach selects a sparse subset of relevant features. The results of feature selection are independent of the regression model and apply universally. Unlike wrapper or embedded methods of feature selection, the proposed method produces interpretable results. Since there is no interference from the regression model, the results of feature selection can be used for domain analysis.

- The proposed method preserves multi-way data structure, which is advantageous in two ways. Firstly, it considers correlations between all modes, contrarily to the unfolded method. Secondly, it allows a more efficient formulation of the optimization problem.

The proposed modification of QPFS is applied to the problem of hand trajectory prediction. The quality of simple linear regression based on QPFS-selected features is comparable with the quality PLS regression. We observed that multi-way QPFS produced similar results to those of QPFS in terms of regression quality and did it much more efficiently due to exploitation of multi-way structure of the data.

The proposed method can be used to construct predictive models in Brain-Computer Interface design. We believe that our optimization-based feature selection algorithm will be able to provide efficient solution for feature selection problem in other multi-way settings.

Future research. In the course of BCI design, ECoG analysis and motor imagery reconstruction, a promising direction for future research is automatic feature generation. In this paper we have adopted the traditional approach to brain signal analysis based on spectral transforms. However, many powerful algorithms (Bengio et al., 2012) were proven able to compete with feature extraction based on expert knowledge. Convolution neural networks were shown (Lawhern et al., 2018; Walker & Deisenroth, 2015) as accurate as conventional algorithm based on spectral features. 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.

For the proposed algorithm it may be useful to investigate other methods of measuring multi-way similarity between features. As for multi-way feature selection in general, an interesting research direction is to develop an algorithm for simultaneous object and feature selection in the style of CUR-decomposition (Mahoney et al., 2008). Model-free selection of most representative objects might also provide useful insight into the problem domain and help reduce costs for further data representation.

6. Acknowledgements

This work was supported by the Russian Foundation for Basic Research (grant 16-07-01158a) and by the Government of the Russian Federation (Agreement 05.Y09.21.0018).

AppendixA. Derivation of multi-way QPFS

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

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

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

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

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

This problem solves iteratively, via alternate approach. At each step a quadratic program 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:

$$\boldsymbol{\alpha}_1^{(k)} = \underset{\boldsymbol{\alpha} \in \{0,1\}^{nR}}{\arg\min}\ \boldsymbol{\alpha}_1^\intercal \left(\tilde{\mathbf{Q}}_1^{(k-1)} \boldsymbol{\alpha}_1 + \tilde{\mathbf{I}}_1^{(k-1)} \right) + \tilde{\mathbf{B}}_1^{(k-1)} \boldsymbol{\alpha}_1, \tag{A.3}$$

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

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

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

   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 multi-way QPFS.

References

Bengio, Y., Courville, A., & Vincent, P. (2012). Representation learning: A review and new perspectives. arXiv:1206.5538.

Bougrain, L., & Liang, N. (2009). Band-specific features improve finger flexion prediction from ECoG.

Bundy, D. T., Pahwa, M., Szrama, N., & Leuthardt, E. C. (2016). Decoding three-dimensional reaching movements using electrocorticographic signals in humans. *Journal Of Neural Engineering*, *13*, 026021.

Cao, B., He, L., Kong, X., Yu, P. S., Hao, Z., & Ragin, A. B. (2014). Tensor-based multi-view feature selection with applications to brain diseases. *Proceedings. IEEE International Conference on Data Mining*, (pp. 40–49).

Chao, Z., Nagasaka, Y., & Fujii, N. (2010). Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkeys. *Frontiers in Neuroengineering*, *3:3*.

Eliseyev, A., & Aksenova, T. (2016). Penalized multi-way partial least squares for smooth trajectory decoding from lectrocorticographic (ECoG) recording. (p. e0154878). volume 11.

Eliseyev, A., Moro, C., Costecalde, T., Torres, N., Gharbi, S., Mestais, C., Benabid, A. L., & Aksenova, T. (2011). Iterative N-way partial least squares for a binary self-paced brain-computer interface in freely moving animals. *Journal of Neural Engineering*, *8*, 046012.

Katrutsa, A., & Strijov, V. (2017). Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. *Expert Systems with Applications*, *76*.

Kim, T.-K., Wong, S.-F., & Cipolla, R. (2007). Tensor canonical correlation analysis for action classification. *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07*, .

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

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). Eegnet: A compact convolutional network for eeg-based brain-computer interfaces. *arXiv:1611.08024v3*, .

Li, J., & Zhang, L. (2009). Regularized tensor discriminant analysis for single trial eeg classification in bci. *Pattern Recognition Letters*, .

Lotric, M. B., Stefanovska, A., Stajer, D., , & Rovan, V. U. (2000). Spectral components of heart rate variability determined by wavelet analysis. *Physiology Measurements*, *21*, 441–457.

Mahoney, M. W., Maggioni, M., & Drineas, P. (2008). Tensor-cur decompositions for tensor-based data. *SIAM Journal on Matrix Analysis and Applications*, *30*, 957–987.

Shimoda, K., Nagasaka, Y., Z.C., C., & Fujii, N. (2012). Decoding continuous three-dimensional hand trajectories from epidural electrocorticographic signals in japanese macaques. *Journal of Neural Engeneering*, (p. 036015). http://neurotycho.org/food-tracking-task.

Smalter, A., Huan, J., & Lushington, G. (2009). Feature selection in the tensor product feature space. *Proceedings. IEEE International Conference on Data Mining*, (pp. 1004–1009).

Walker, I., & Deisenroth, M. P. (2015). Deep convolutional neural networks for brain computer interface using motor imagery.

Zhao, Q., Zhou, G., Adali, T., & Cichocki, A. (2013). Kernelization of tensor-based models for multiway data analysis. *IEEE Signal Processing Magazine*, *30*, 137–148.