Quadratic Programming Feature Selection for Multicorrelated Signal Decoding with Partial Least Squares

R.V. Isachenko, V.V. Strijov

**Abstract:** This paper is devoted to the dimensionality reduction problem in signal decoding. The challenge of this investigation is the redundancy in the data description. High correlations among measurements lead to correlations in the input space. This study considers the multivariate problem, and the target variable is a vector. In this case, the correlations occur in both the input and target spaces. Dimensionality reduction and feature selection are used to build a simple and stable model.

The partial least squares (PLS) regression is used as the base model for the dimensionality reduction. The model projects the input and target data into the joint latent space and maximizes the covariances between the projections. To obtain the sparse model, feature selection is applied. The majority of feature selection methods ignore the dependencies in the target space. The study suggests a novel approach to feature selection using a multivariate regression. The proposed approach extends the ideas of the quadratic programming feature selection (QPFS) algorithm. The QPFS algorithm selects the noncorrelated features that are relevant to the targets. The proposed methods take into account the dependencies in the target space and select the features that are informative to all targets jointly.

The computational experiment was carried out using the electrocorticogram (ECOG) dataset. The proposed algorithms were compared using different criteria such as their stability and predictive performance. The algorithms give significantly better results compared to the baseline strategy. The QPFS approach was compared with the partial least squares (PLS) regression. The best result is obtained by a combination of the QPFS and PLS algorithms.

**Keywords**: partial least squares, quadratic programming feature selection, signal decoding

Introduction

The initial data in the fields of chemometrics [1, 2] and signal decoding [3, 4] are high-dimensional and extremely redundant. The models that are built on such data are unstable. In addition, the redundant data description requires excessive computations, which lead to extended analysis times. To overcome this problem, dimensionality reduction [5, 6] and feature selection [7, 8] methods are used for high-dimensional data modeling.

The partial least squares (PLS) is a widely used algorithm for dimensionality reduction [9-12]. The PLS finds the optimal combinations of the initial features and uses these combinations as the model features. The algorithm projects the features and the targets onto the joint latent space and maximizes the covariances between the projected vectors. It allows researchers to save the information about the initial input and target matrices and find their relations. The dimensionality of the latent space is much less than the size of the initial data description. It leads to a stable linear model built on a small number of features. An overview of the advances in the PLS regression is given in [13, 14]. For this model, we obtain the linear model with a small latent dimension. However, the final model uses the whole range of the initial features, and it does not allow for the removal of useless features.

Feature selection is a special case of dimensionality reduction when the latent representation is a subset of the initial data description. Here, the models are built on the subsets of the features. One of the approaches to feature selection is to maximize the feature relevances and minimize the pairwise feature redundancy. This approach was recently proposed and investigated in [15, 16]. Quadratic programming feature selection (QPFS) [17] uses this approach to construct the optimization problem. It was shown in [18] that the QPFS algorithm outperforms many existing feature selection methods for the univariate regression problem. The QPFS algorithm introduces two functions: and . estimates the redundancy between features, and contains the relevances between each feature and the target vector. QPFS minimizes the Sim function and maximizes the Rel function simultaneously. The algorithm solves the following optimization problem:

(1)

Here, the columns of matrix  are the features, and  is the target vector. The matrix  entries measure the pairwise similarities between features. The vector expresses the similarities between each feature and the target vector. The normalized vector  shows the importance of each feature. Function (1) penalizes the dependent features using the Sim function and promotes the features that are relevant to the target using the Rel function. The parameter  controls the trade-off between Sim and Rel. To measure the similarity, the authors of [17] use the absolute value of the sample correlation coefficient between pairs of features for the Sim function, and between the features and the target vector for the Rel function.

Paper [19] proposes a multiway version of the QPFS algorithm for tensor ECoG-based data. It was shown that QPFS is an appropriate feature selection method for the signal decoding problem. We consider the multivariate problem where the dependent variable is a vector. It leads to correlations in the model targets. In this situation, feature selection algorithms do not take into account these dependencies. Hence, the selected feature subset is not optimal in terms of its prediction. We propose methods that take into account the dependencies in both the input and target spaces. It allows us to form a stable sparse model. We refer to the original QPFS algorithm as our baseline for the computational experiment.

The main drawback of the QPFS algorithm is its computational costs. However, the original paper [17] suggests a way to solve the quadratic problem (1) efficiently. Additionally, in [20], the sequential minimal optimization framework is proposed for solving (1).

The experiments were carried out using the ECoG dataset [21]. We compared the proposed methods for multivariate feature selection with the baseline strategy and the PLS algorithm. The stability of the proposed methods was investigated by measuring how the feature selection solution changes with data bootstrapping. The proposed algorithms outperform the baseline algorithm given the same number of features. The combination of the feature selection procedure and the PLS algorithm gives the best performance.

The main contributions of this paper are as follows:

* addressing the dimensionality reduction problem for high-dimensional data,
* proposing new feature selection methods for multivariate regressions with the analysis of the input and target spaces structures,
* comparing the proposed methods using a real ECoG dataset, and showing that the proposed methods give better feature subsets than the baseline method.

Multivariate regression

The goal is to forecast a dependent variable with targets from an independent input object with features. We assume that there is a linear dependence between the object and the target variable as

, (2)

where is the matrix of the model parameters, and is a residual vector. One can find the matrix of the model parameters  given a dataset using the following formula, where is a design matrix and is a target matrix:

The columns  of  correspond to the object features, and the columns  of  correspond to the targets.

The optimal parameters are determined by the minimization of an error function. We define the quadratic loss function as follows:

(3)

The solution of (3) is given by

The linear dependent columns of  lead to an unstable solution for the optimization problem (3). If there is a vector such that , then adding  to any column of  does not change the value of the loss function . In this case, the matrix  is close to singular and is not invertible. To avoid strong linear dependence, dimensionality reduction and feature selection are used.

Feature selection

The feature selection goal is to find the boolean vector  in which the components indicate whether the feature is selected. To obtain the optimal vector  among all possible options, we introduce the feature selection error function . We state the feature selection problem as follows:

(4)

The goal of feature selection is to construct the appropriate function . The particular examples for the considered feature selection algorithms are given below and summarized in Table 1.

Problem (4) is hard to solve due to the discrete binary domain . We relax problem (4) to the continuous domain . The relaxed feature selection problem is

(5)

Here, vector ’s entries are the normalized feature importances. First, we solve problem (5) to obtain the feature importances . Then, the solution of (4) is recovered by thresholding as follows:

  is a hyperparameter that is defined manually or chosen by cross-validation.

Once the solution  of (4) is known, problem (3) becomes

where subscript  indicates the sub matrix with the columns in which the components of  equal 1.

## Quadratic Programming Feature Selection

Paper [18] shows that QPFS outperforms many existing feature selection algorithms using different quality criteria. The QPFS algorithm selects the noncorrelated features that are relevant to the target vector  for the linear regression problem where  as follows:

The authors of the original QPFS paper [17] suggested the following way to select  for (1) and make and have the same impacts:

The QPFS parameters are defined as follows:

(6)

Here,  is the absolute value of the sample Pearson correlation coefficient:

Other ways to define and are considered in [18].

Problem (1) is convex if the matrix  is positive semidefinite. In general, this is not always true. To satisfy this condition, the matrix  spectrum is shifted and matrix  is replaced by , where is the minimum eigenvalue of .

## Multivariate QPFS

Here, we propose the algorithms for feature selection in the multivariate case. If the target space is multidimensional, it is prone to redundancy and correlations between the targets. In this section, we propose the algorithms that take into account the dependencies in both the input and target spaces.

## Relevance aggregation (RelAgg).

In (19), in order to apply the QPFS algorithm to the multivariate case (), feature relevances are aggregated through all components. The term is still the same, and matrix  is defined by (6). The vector is aggregated across all targets and is defined as

The drawback of this approach is its insensitivity to the dependencies in the columns of . Observe the following example:

We have 3 features and targets, where the first targets are identical. The pairwise features similarities are given by matrix . Matrix ’s entries give the pairwise features relevances of the targets. Vector is obtained by the summation of matrix  over the columns:

(7)

We would like to select only two features. For such a configuration, the best feature subset is . Feature  predicts the second target  and the combination of features  predicts the first component. The QPFS algorithm for  gives the solution . It coincides with our knowledge. However, if we add the collinear columns to matrix  and increase  to 5, the QPFS solution will be . Here, we lose the relevant feature  and select the redundant feature . The following subsections propose extensions to the QPFS algorithm that overcome the challenge of this example.

## Symmetric importances (SymImp).

To take into account the dependencies in the columns of matrix , we extend the QPFS function (1) to the multivariate case. We add the term  and modify the term as follows:

(8)

We determine the entries of matrices , , and in the following way:

Vector  shows the features’ importances, while is a vector of the targets’ importances. The correlated targets will be penalized by and have lower importances.

The coefficients , , and control the influence of each term on function (8) and satisfy the following conditions:

**Proposition 1.** *The balance between the terms , , and for the problem is achieved by the following coefficients:*

(9)

The desired values of , , and are given by solving of the following equations:

Here, the mean values , , and of the corresponding matrices , , and are the mean values of the terms , , and .

To investigate the impact of  on function (8), we balance the terms and by fixing the proportion between  and :

We apply the proposed algorithm to the discussed example (7). The given matrix  corresponds to matrix . We additionally define matrix  by setting and all others entries to one. Figure 1 shows the importances of features  and targets  with respect to . If  is small, the impacts of all targets are almost identical and feature  dominates feature . When  becomes larger than , the importance  of target  increases along with the importance of feature .

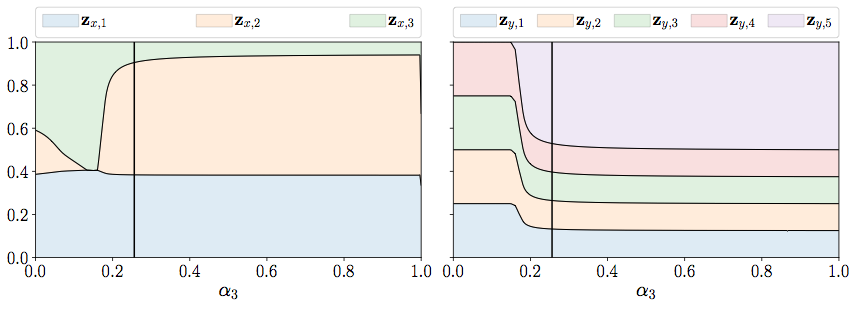


Figure 1. Feature importances and w.r.t.  for the considered example

## Minimax QPFS (MinMax).

Function (8) is symmetric with respect to  and . It penalizes the features that are correlated and irrelevant to the targets. In addition, it penalizes the targets that are correlated and are not sufficiently explained by the features. It leads to small importances for the targets that are weakly correlated with the features and large importances for the targets that are strongly correlated with the features. This result contradicts the intuition. Our goal is to predict all targets, especially those that are difficult to explain, using the selected relevant and noncorrelated features. We express this as two related problems:

(11), (12)

The difference between (11) and (12) is the sign of Rel. In the input space, the nonrelevant components should have smaller importances. Meanwhile, the targets that are not relevant to the features should have larger importances. Problems (11) and (12) are merged into the joint min-max or max-min formulation

(13)

where

**Theorem 1.** *For positive definite matrices and* , *the max-min and min-max problems (13) have the same optimal value.*

We denote the following:

The sets and are compact and convex. The function is a continuous function. If and are positive definite matrices, function  is convex-concave. I.e., is convex for a fixed , and is concave for a fixed . In this case, Neumann’s minimax theorem states that

To solve the min-max problem (13), we fix some . For a fixed vector , we solve the problem

(14)

The Lagrangian for this problem is

Here, the Lagrange multipliers that correspond to the inequality constraints are restricted to being nonnegative. The dual problem is

(15)

The strong duality holds for quadratic problem (14) with the positive definite matrices  and . Therefore, the optimal value for (14) equals the optimal value for (15). It allows us to solve the problem

(16)

instead of (13).

By setting the gradient of the Lagrangian equal to zero, we obtain an optimal value for :

(17)

The dual function is equal to

(18)

It represents the quadratic problem (16) with  variables.

## Asymmetric Importance (AsymImp)

The natural way to overcome the problem of the SymImp strategy is to add penalties for targets that are correlated with features. We add the term  to the term as follows:

(19)

**Proposition 2.** *Let vector equal*

*Then, the importance coefficients for vector  will be nonnegative in  for problem (19).*

The proposition follows from the fact that

where and .

Hence, function (19) discourages the features that are relevant to the targets and encourages the targets that are not sufficiently correlated with the features.

**Proposition 3.** *The balance between the terms , , and for the problem is achieved by the following coefficients:*

The desired values of , , and are given by the solutions to the following equations:

(20), (21), (22)

Here, we balance with the first term of  using (21) and with the full  using (22).

**Proposition 4.** *For the case of , the proposed functions (8), (13), and (19) coincide with the original QPFS algorithm (1).*

If is equal to 1, then is a scalar, , and . It reduces problems (8), (13), and (19) to

Setting represents the original QPFS problem (1).

Table 1 shows the core ideas and error functions for each method and summarizes all the proposed strategies for multivariate feature selection. RelAgg is the baseline strategy, and it does not consider the target space correlations. SymImp penalizes the pairwise target correlations. MinMax are more sensitive to the targets that are difficult to predict. The AsymImp strategy adds the term to the SymImp function to make the features and targets have asymmetric influences.

*——————-*

Table 1: Overview of the proposed multivariate QPFS algorithms

*——————-*

## Dimensionality reduction

To eliminate the linear dependence and reduce the dimensionality of the input space, principal components analysis (PCA) is a widely used algorithm. The main disadvantage of the PCA method is that it is insensitivity to the interrelation between the features and the targets. The partial least squares algorithm projects the design matrix  and the target matrix  to the latent space with low dimensionality (). The PLS algorithm finds the latent space matrices that best describe the original matrices  and .

The design matrix  and the target matrix  are projected into the latent space in the following way:

(23), (24)

where and are the scores matrices in the latent space, and are the loading matrices, and are residual matrices. The PLS maximizes the linear relation between the columns of matrices  and as

We use the PLS algorithm as the dimensionality reduction algorithm in this research.

To obtain the model prediction and find the model parameters, we multiply both sides of (23) by matrix . Since the residual matrix ’s rows are orthogonal to the columns of , we have

The linear transformation between objects in the input and latent spaces is the following:

(25)

The matrix of the model parameters (2) could be found from equations (24) and (25) as

(26)

Thus, the model parameters (2) are equal to

The final model (26) is a low-dimensional linear model in the latent space. It reduces the data redundancy and increases the model stability.

Experiment

To evaluate the selected feature subset, we introduce criteria that estimate the quality of feature selection. We measure the multicorrelation using the mean value of multiple the correlation coefficient as follows:

This coefficient lies between 0 and 1. A bigger means that we have a better feature subset.

The model stability is given by the logarithmic ratio between the minimum eigenvalue  and maximum eigenvalue  of matrix :

A smaller stability value indicates less multicollinearity in matrix .

The scaled Root Mean Squared Error (sRMSE) shows the quality of the model prediction. We estimate the sRMSE using the training and testing data.

Here, is the model prediction and is the constant prediction obtained by averaging the targets across all objects. The error on the testing set should be as small as possible.

The Bayesian Information Criteria (BIC) incorporates a trade-off between the prediction quality and the size of selected subset :

.

A smaller value of BIC means a better feature subset.

## Data

We conducted a computational experiment with the ECoG data from the NeuroTycho project [21]. The ECoG data consist of brain voltage signals recorded over 32 channels. The goal is to predict 3D hand positions in subsequent moments given the input signal. The initial voltage signals are transformed to the spatial-temporal representation using the wavelet transformation with the Morlet mother wavelet. The procedure of extracting the feature representation from the raw data is described in detail in [22, 23]. The data and feature description at each time moment have dimensions of size 32 (channels) 27 (frequencies) = 864. Each object is the representation of the local historical time segment with duration . The time step between objects is . The final matrices are and , where is the number of timestamps that we predict. We split our data into training and testing parts with the ratio 0.67. Examples of the initial brain signals and the corresponding hand movements are shown in Figure 2.

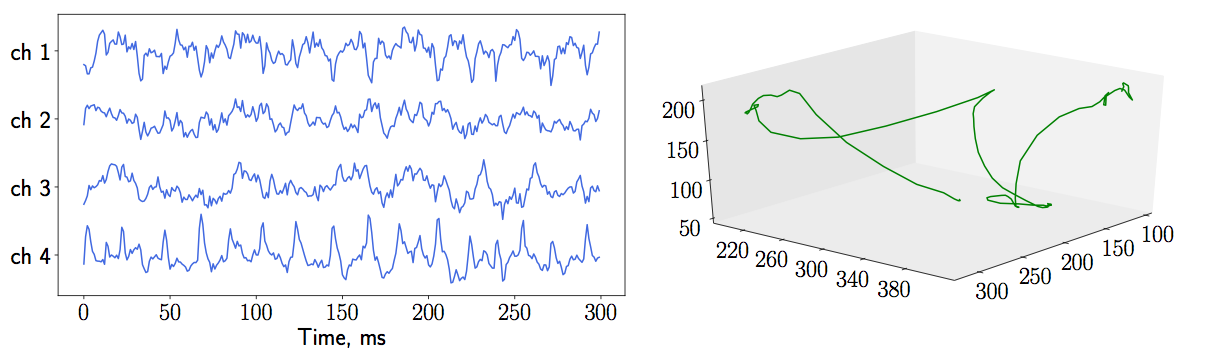


Figure 2. Brain signals (left plot) and 3D hand coordinates (right plot)

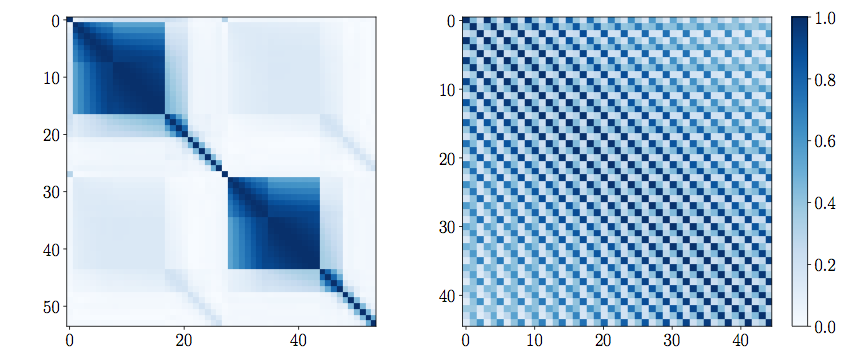
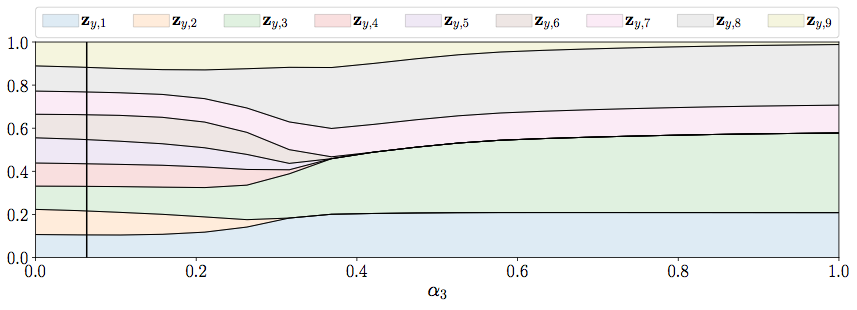
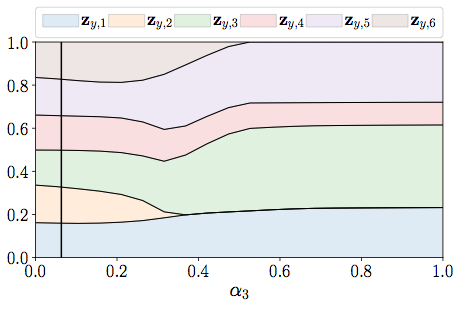
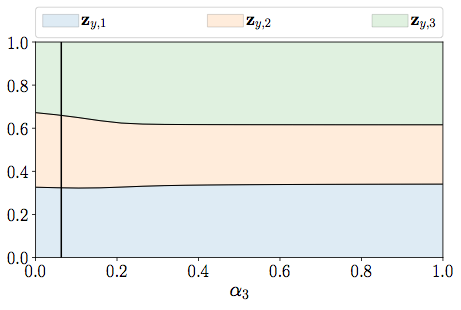
Figure 3 shows the dependencies in matrices  and  for the ECoG data. The frequencies in the matrix  are highly correlated. In the target matrix , the correlations between axes are not significant in comparison with the correlations between consequent moments and these correlations decay with time.

Figure 3. Correlation matrices for and

## Results

We apply the QPFS algorithm with the SymImp strategy for different values of  according to formula (10). The dependencies between target importances  with respect to  for different values of  are shown in Figure 4. The target importances are almost the same for the predicted wrist coordinates with only one timestamp , which reflects the independence between the , , and coordinates. For and , the importances of some targets become zero when  increases. The vertical lines correspond to the optimal value of  obtained by (9). The target importances  for this value of  are similar. Thus, the algorithm does not distinguish the targets for .

*Figure 4. Target importances zy with respect to α3 for QPFS with Symmetric Importance*

We compare the proposed strategies of the multivariate QPFS that are given in Table 1 for the ECoG dataset. First, we apply all the methods to obtain the feature importances. Then, we fit a linear model with an increasing number of features. For each method, the features are sorted by their obtained importances. We show how the described quality criteria change with the increasing feature set size. Figure 5 illustrates the results for the prediction of timestamps. Here, the feature importance threshold is represented by colored ticks. These thresholds are larger for the proposed methods in comparison to the baseline RelAgg strategy. The SymImp strategy has the largest threshold, and it does not allow one to obtain a small feature subset. However, this strategy shows the best performance in terms of the sRMSE using the testing data. The second performance measure is AsymImp. All proposed algorithms give smaller test errors compared to the RelAgg strategy. The stability is also increased for the proposed algorithms. Here, we consider the AsymImp strategy as the best in terms of the predictive quality and the size of selected feature subset.

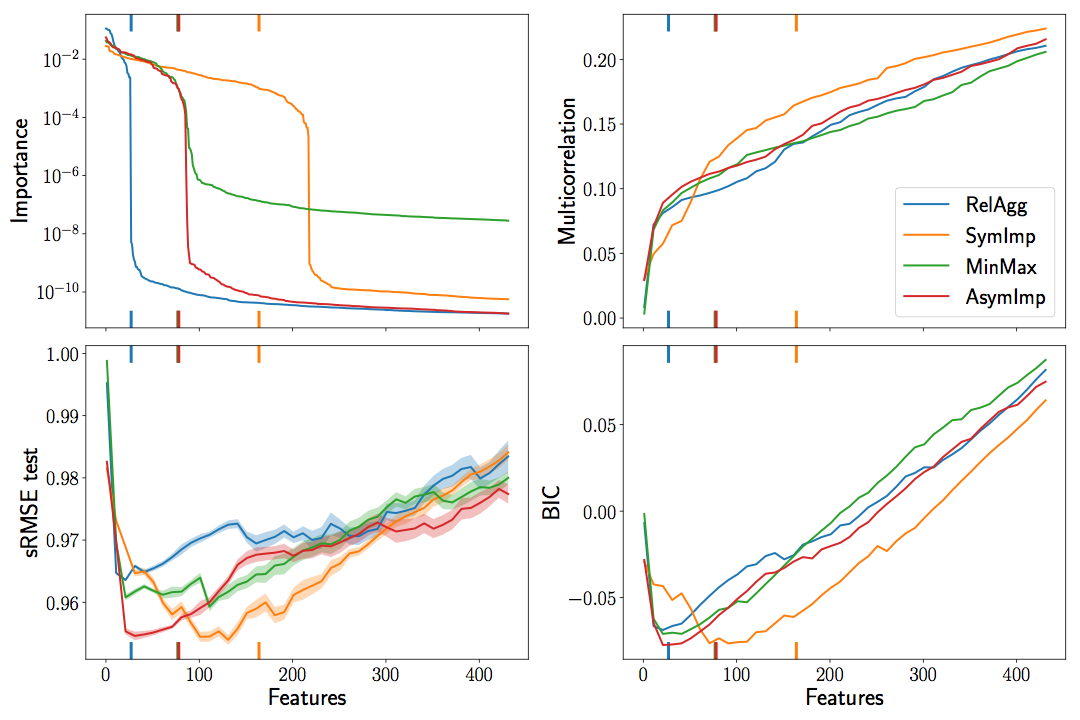


Figure 5. Feature selection algorithms evaluation for the ECoG data and the prediction of timestamps

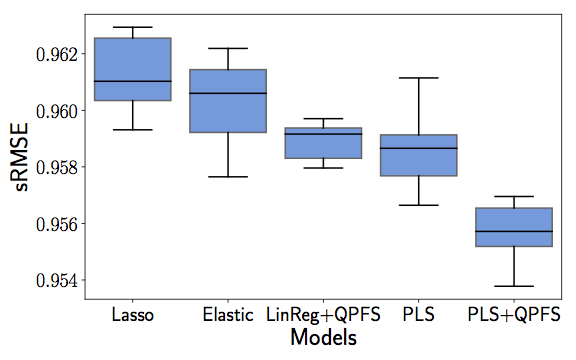
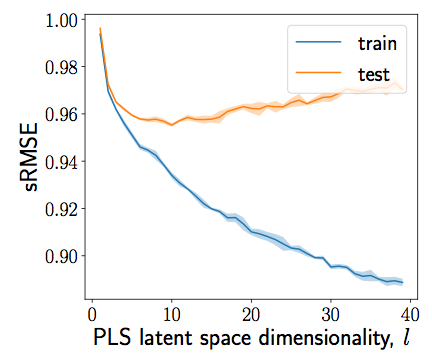
To compare the structure of the selected feature subsets and investigate the stability of the selection procedure, we use the bootstrap approach. First, the bootstrap data are generated. Then, we solve the feature selection problem for each pair of design and target matrices. The obtained feature importances are compared. We calculate the average pairwise Spearman correlation coefficient and the distance as the measures of the algorithm’s stability. Table 2 shows the average error, the size of the subset and the described statistics for each method. The error was calculated by fitting the linear model using the features with the largest importances. AsymImp gives the least error when using the test data. The size of the selected feature subset is overestimated using the threshold . The value of  could be cross-validated to get the optimal threshold and feature subset size.

*——————-*

Table 2: The stability of the selected feature subset

*——————-*

We fit the PLS regression model to the data to compare the dimensionality reduction and feature selection. Figure 6 demonstrates the scaled RMSE on the training and testing data with respect to the dimensionality of the latent space . The testing error reaches its minimum at . The PLS regression is more a flexible approach compared to the linear model built on the subset of features. It results in the smallest error, but the model is not sparse.

Figure 7 compares 3 models: the linear regression and the PLS regression built on 100 features given by QPFS and the PLS regression with all features. We do not include the linear regression with all features because its results are close to the constant prediction. It also provides the result for the Lasso and Elastic Net algorithms that are widely used for feature selection. We use the AsymImp strategy for QPFS in this experiment. Here, the PLS regression is significantly better than the linear regression with the QPFS features. It means that the latter model is not flexible enough. However, the best result is obtained by the PLS regression model combined with the QPFS features. This model is sparse since it uses only 100 QPFS features. The ability of the PLS model to find the optimal latent data representation improves the model performance.

*Figure 6: Test scaled RMSE for PLS Figure 7: sRMSE box plots for different regression models*

Conclusion

The study investigates the problem of signal decoding in which the data are highly redundant. To build a stable, adequate model, we reduced the dimensionality of the problem using the dependencies in both the input and target spaces. The PLS regression is considered as a linear model for dimensionality reduction. The quadratic programming approach is investigated as a feature selection algorithm. The algorithm solves feature selection in a single quadratic programming optimization problem. The multivariate extensions for the QPFS algorithms are proposed. The resulting feature subset includes noncorrelated features that are relevant to the most difficult targets.

The computational experiments were carried out using the ECoG data. The resulting model predicts the limb position of an exoskeleton using brain signals. The proposed algorithm outperforms the baseline algorithm and reduces the problem dimension significantly. The combination of feature selection for sparsifying the model and the dimensionality reduction for increasing the model stability give the best result.