DLCP 2024

Nonlinear relevance estimation of multicollinear features for reducing the input dimensionality of optical spectroscopy inverse problem

Nickolay O. Shchurov[1, 2],
Igor V. Isaev[2, 3], Olga E. Sarmanova[1, 2],
Sergey A. Burikov[1, 2], Tatiana A. Dolenko[1, 2],
Kirill A. Laptinskiy[2], Sergey A. Dolenko[2]

[1] Faculty of Physics, Moscow State University, Moscow, Russia

[2] D.V. Skobeltsyn Institute of Nuclear Physics, M.V. Lomonosov Moscow State University, Moscow, Russia

[3] Kotelnikov Institute of Radioengineering and Electronics, Russian Academy of Sciences, Moscow, Russia

The research was carried out at the expense of the grant of the Russian Science Foundation https://rscf.ru/en/project/24-11-00266/.

Determination of ion concentrations in solutions

Necessity:

- Monitoring in the fields of ecology and industry
- Assessment of the composition of technical, waste and mineral waters

Disadvantages of traditional chemical analytical methods:

- Time-consuming
- Requires:
  - Sample preparation
  - Qualified personnel
  - Expensive reagents

Optical spectroscopy methods solve these problems,

and are also non-contact

Properties of the inverse problem of spectroscopy

- Nonlinear
- Has neither analytical nor direct numerical solution
- Multiparameter
- High input dimension
- High multicollinearity

Machine learning methods are used to process such data

Data Description

- Raman and absorption spectra were obtained experimentally
- The studied solutions contained:
  - 1 to 6 salts
  - 2 to 6 ions
- Concentration range: 0-0.14 M in increments of 0.01 M.

- For each sample the following were measured:
  - Optical absorption spectrum,
  - Raman spectrum
  - pH value.

Data Description

- Number of patterns in the original dataset: 3 806
- Dividing into sets:
  - ✓ Training 70% 2 656 patterns
  - ✓ Validation 20% 750 patterns
  - ✓ Test 10% 400 patterns
- Data dimension:
  - ✓ Input 2 048+811 features
  - ✓ Output 6 features

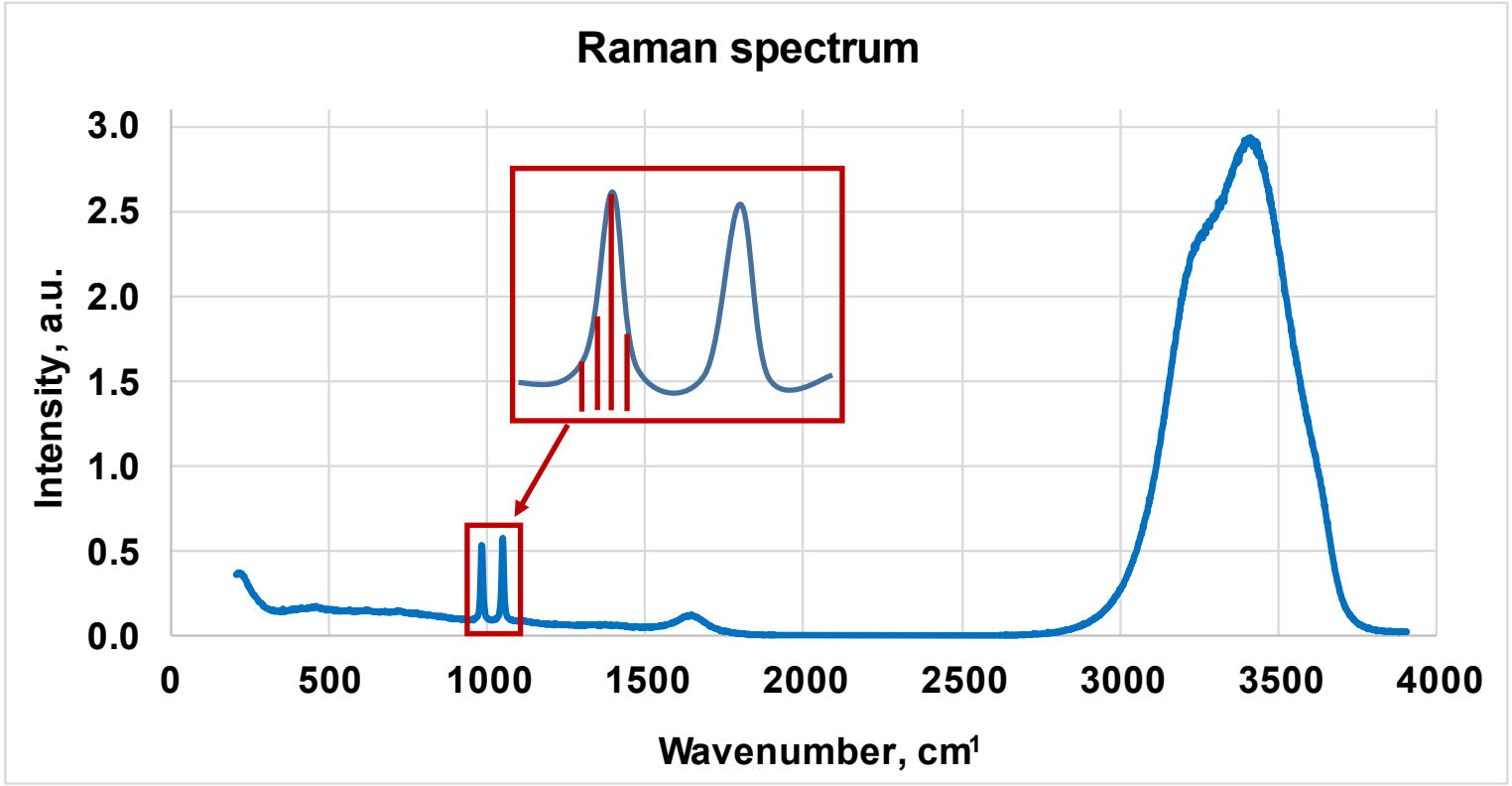
Multicollinearity

Multicollinearity – the presence of a linear relationship between features

The features in this work are the **values of intensity** in the spectrum channels

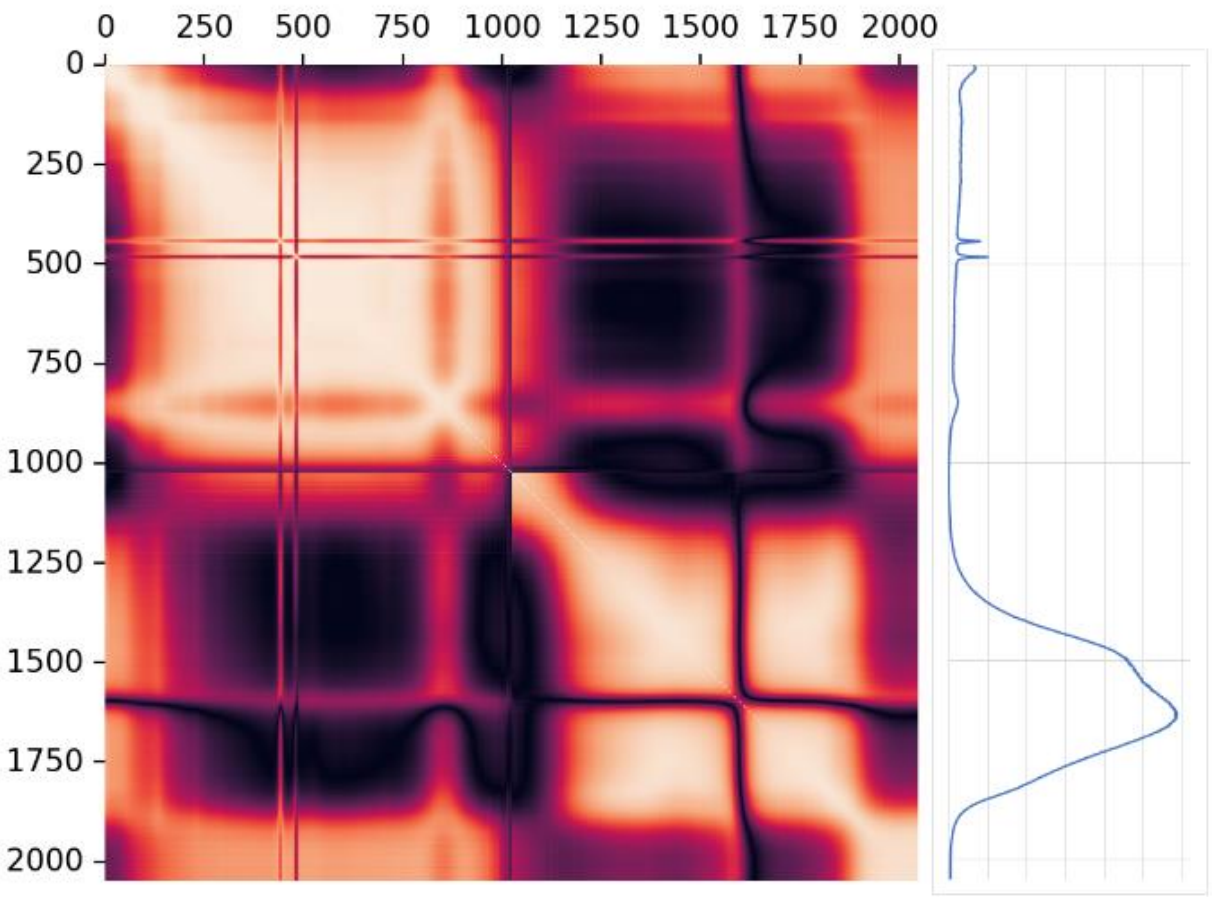
Reasons for multicollinearity of spectroscopic data:

1. Spectral bands are several spectrum channels wide
2. Close spectrum channels carry similar information

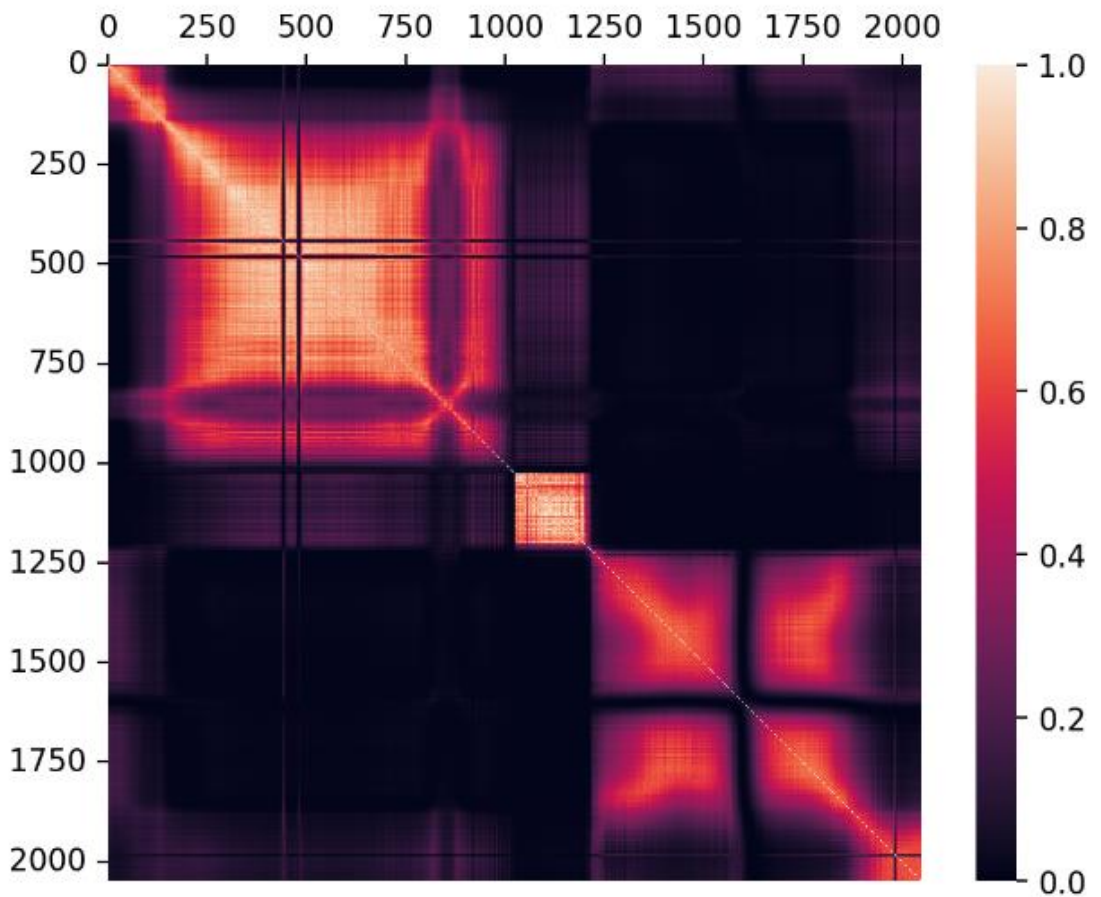


Heat maps

Cross correlation



Cross entropy (with normalization)



Feature Selection

Feature selection is the procedure of removing features before running machine learning.

Reducing the number of input features is necessary for several reasons:

- The computational cost is reduced
- Reduced minimum training set size requirements
- Solution accuracy may improve
- The different significance of features is taken into account

Goal: select relevant features and eliminate redundant ones

Methods for selecting essential features

1. **Wrappers –** methods, based on repeated solution of the problem using different subsets of features
   - Pros: allow you to detect possible relationships between variables
   - Cons: very computationally complex when the number of features is large
2. **Filters** – assessment of significance based on initial data
   - Pros : computationally simple
   - Cons: relationship between features are usually not taken into account
3. **Built-in** – the selection procedure is built into the algorithm for solving the problem
   - Pros: the same algorithm is used for selection as for solving the problem
   - Cons: selection results vary from run to run

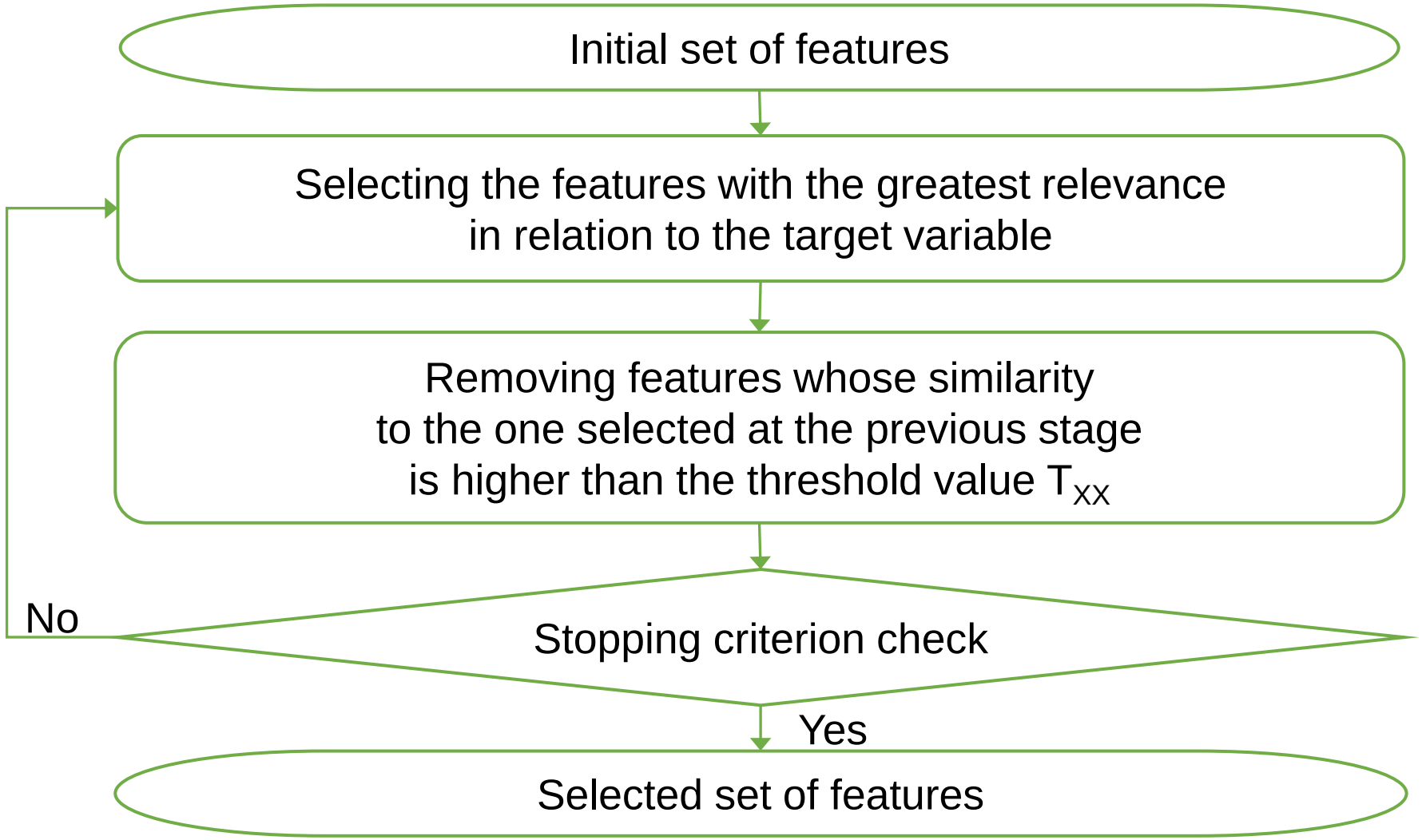
Goal of the work

- Research of filter type selection method,
  taking into account the correlation between input features.

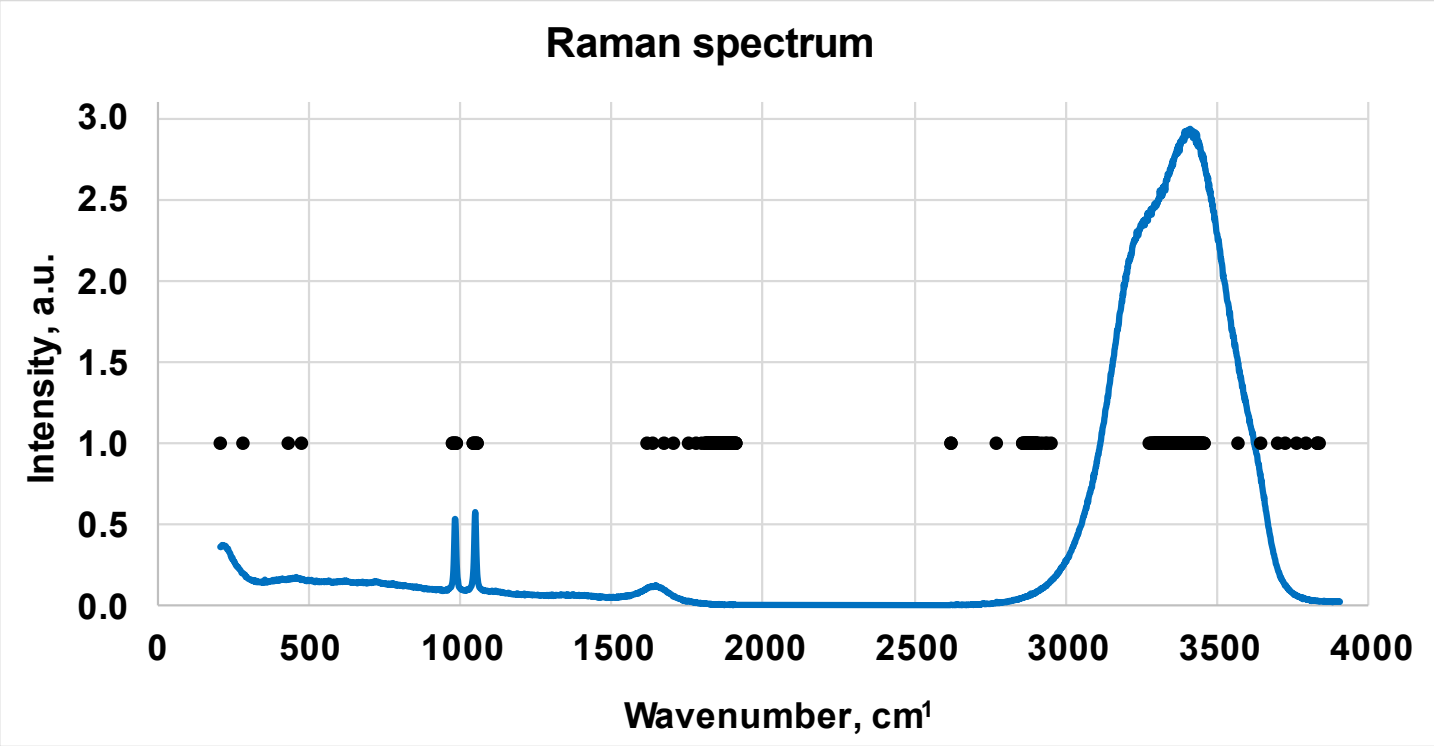
Tasks

- Determine the optimal parameters of the algorithm
- Compare the accuracy of neural networks trained:
  1. On the full set and on its subsets obtained using feature selection
  2. Using different parameters of the feature selection method
  3. Using the combination of Raman and absorption spectrum

Iterative feature selection algorithm (IFS)



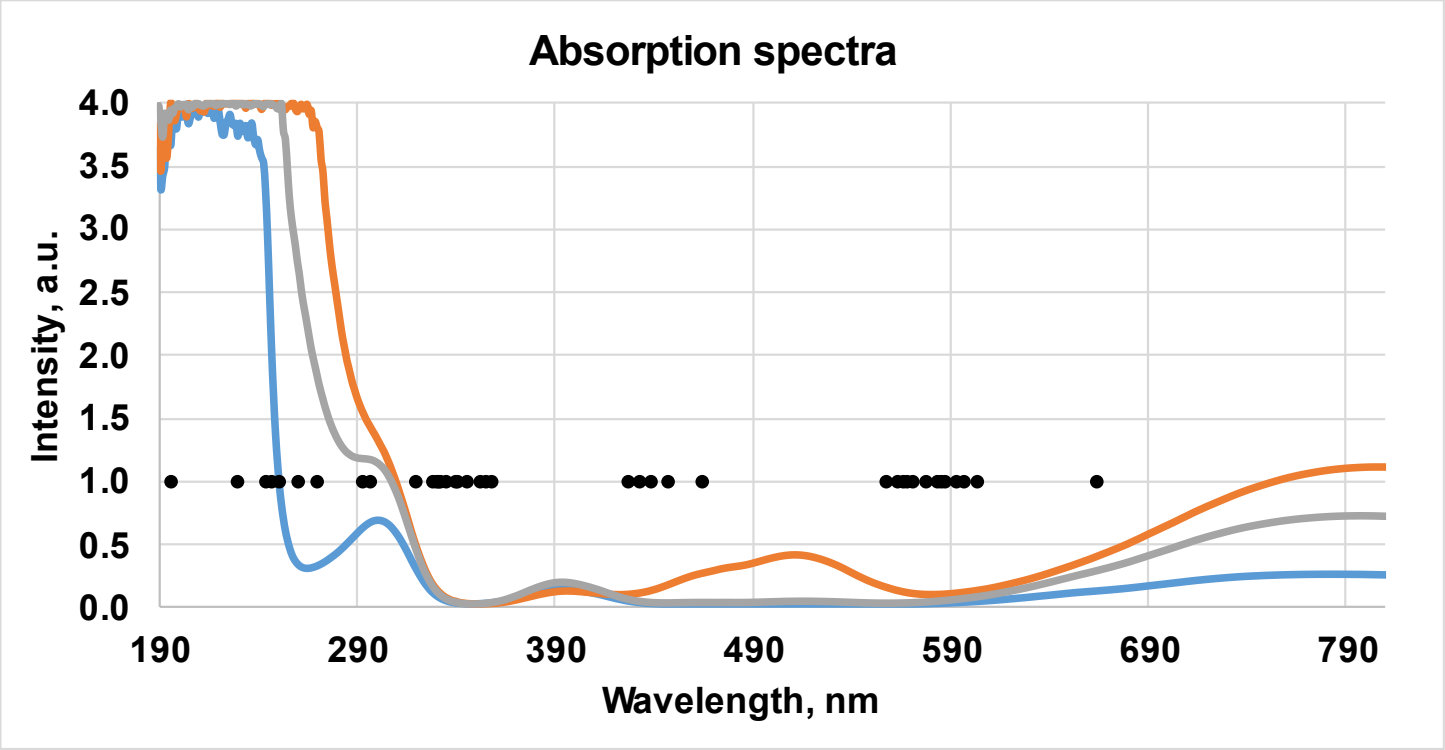
Selected features



The selected features mainly relate to:

- Valence band of water
- Characteristic bands of ions

Selected features

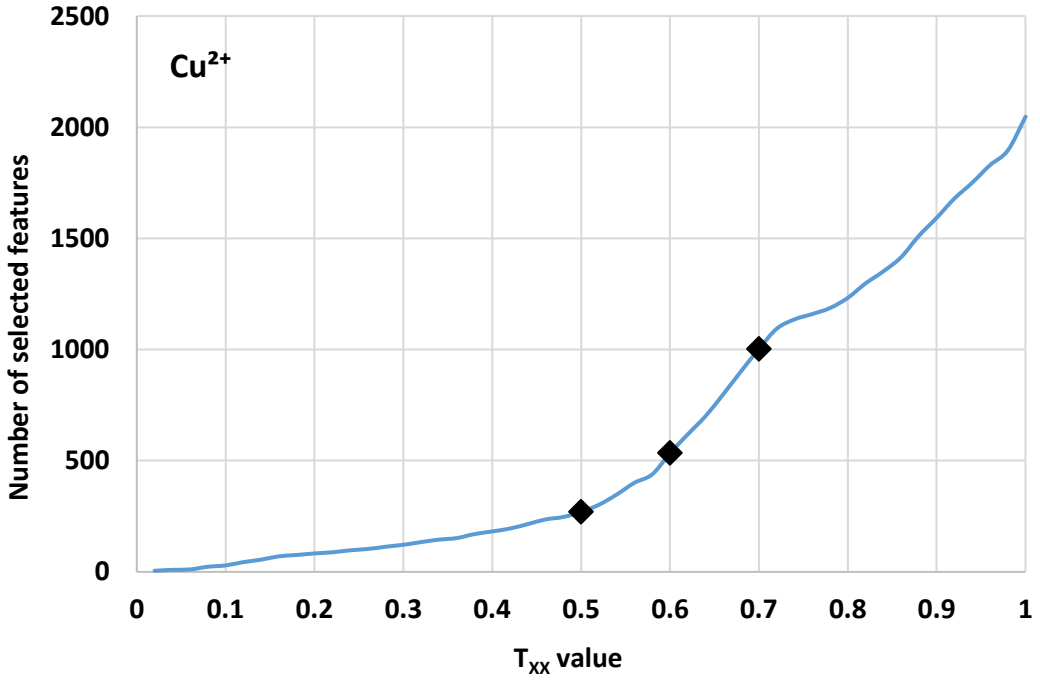


- The selected features mainly belong to the edges of the spectral bands, and not to their maxima.
- This phenomenon can be explained by the fact that spectral lines overlap, and the areas with the least overlap of bands are the most sensitive

Neural network parameters

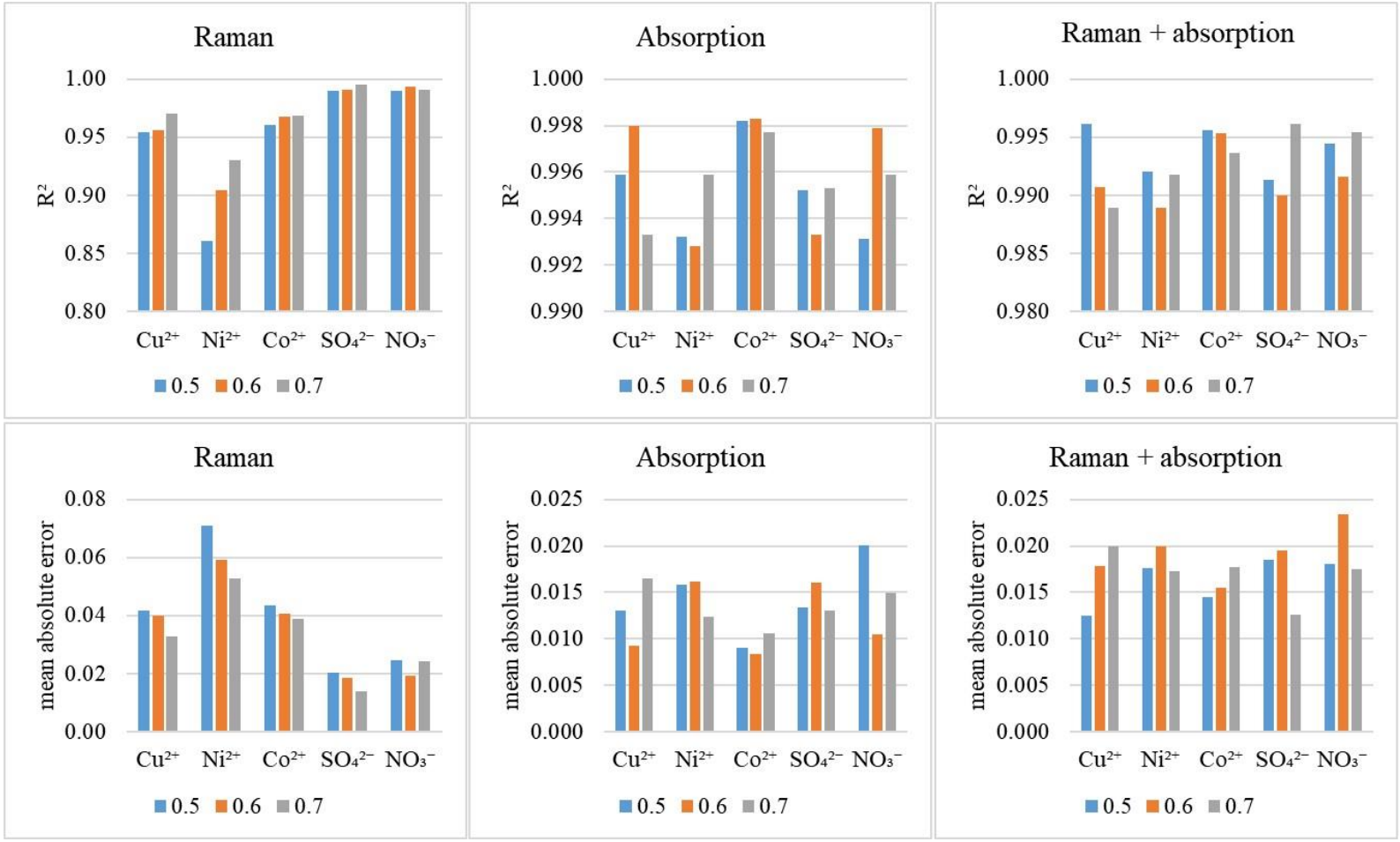
- Architecture: multilayer perceptron with 1 hidden layer.
- Number of neurons in the hidden layer: 32.
- Activation function:
  Hidden layer: sigmoid
  Output layer: linear
- Early stop:
  800 epochs with no improvement on validation set
- Each neural network was trained 5 times with different initial values of the weights.
  Statistical indicators of the results of using 5 networks were averaged.

Threshold selection



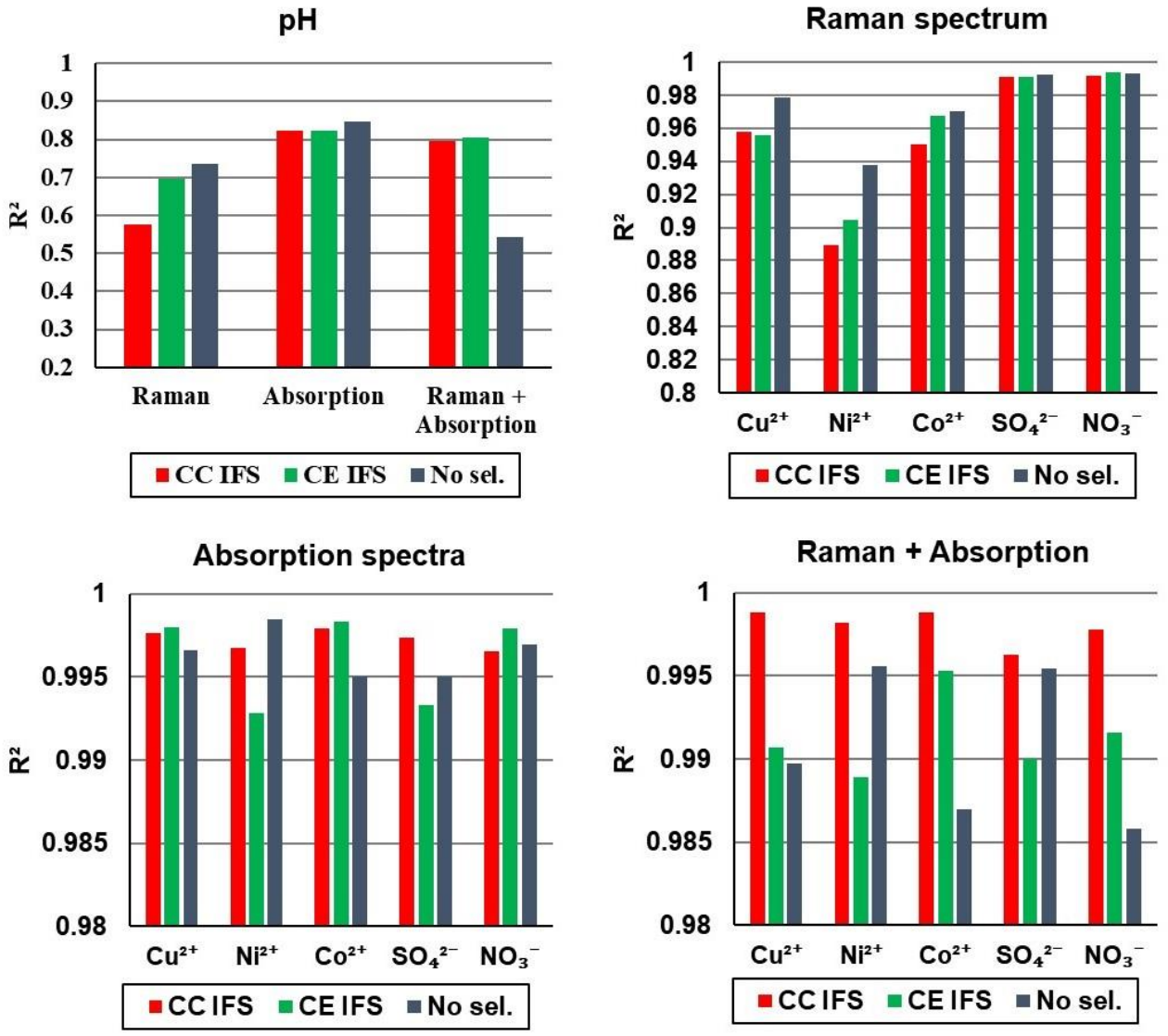
- As threshold values T_{XX} for further consideration, four values were selected from the inflection area of the graph
- The choice of the optimal threshold value is based on the results of solving the problem on selected sets of features

Determination coefficient R^2 of neural network solution using IFS algorithm depending on the threshold value T_{xx}

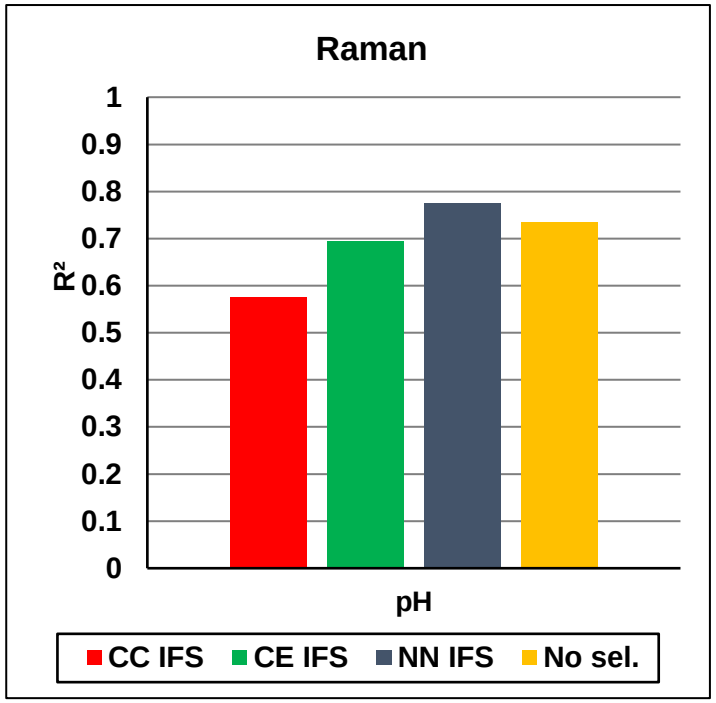


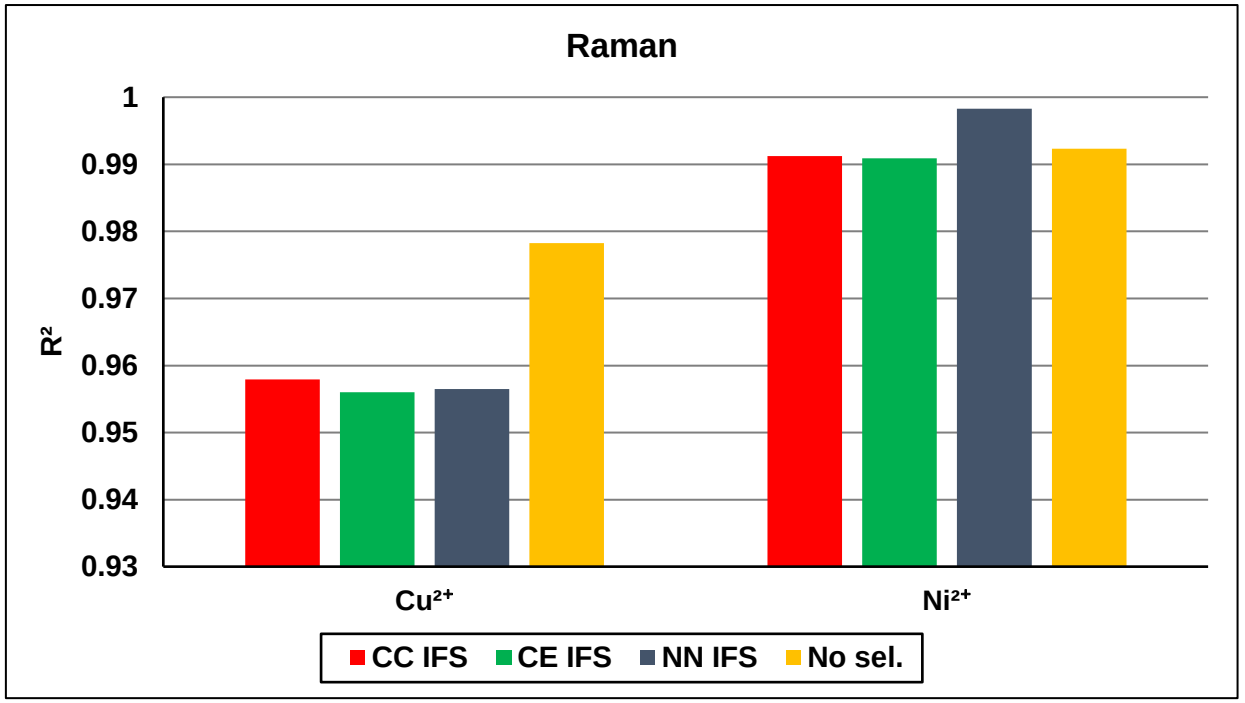
The threshold chosen as optimal: T_{xx}= 0.6

R^2 of neural network solution on various data



Using different relevance metrics





Conclusions

IFS method allows to:

- Significantly reduce the dimensionality of the input data while maintaining the quality of the solution
- Achieve a better solution when combining different data, especially if the dimension of this data is large

Weight analysis as a nonlinear relationship metric allows to get better results for some ions

Thank you for your attention!