ROBUST CLASSIFICATION OF SPECTROSCOPIC DATA IN AGRI-FOOD: FIRST ANALYSIS ON THE STABILITY OF RESULTS

Andrea Cappozzo[1], Ludovic Duponchel[2], Francesca Greselin[3] and Brendan Murphy[4]

[1] Department of Mathematics, Politecnico di Milano, (andrea.cappozzo@polimi.it)

[2] LASIR Lab, University of Lille, (ludovic.duponchel@univ-lille.fr)

[3] Department of Statistics and Quantitative Methods, University of Milano Bicocca, (francesca.greselin@unimib.it)

[4] School of Mathematics and Statistics, University College Dublin, (brendan.murphy@ucd.ie)

ABSTRACT: We investigate here the stability of the obtained results of a variable selection method recently introduced in the literature, and embedded into a model-based classification framework. It is applied to chemometric data, with the purpose of selecting a few wavenumbers (of the order of tens) among the thousands measured ones, to build a (robust) decision rule for classification. The robust nature of the method safeguards it from potential label noise and outliers, which are particularly dangerous in the field of food-authenticity studies. As a by-product of the learning process, samples are grouped into similar classes, and anomalous samples are also singled out. Our first results show that there is some variability around a common pattern in the obtained selection.

KEYWORDS: Variable selection, Robust classification, Label noise, Outlier detection, Near infrared spectroscopy, Mid infrared spectroscopy, Agri-food.

1 Introduction

Nowadays, many challenging classification problems, arising from scientific domains such as chemometrics, computer vision, engineering, and genetics, among others, have to deal with hundreds or thousands of variables on each sample. Many contributions in the literature show that inferential methods benefit greatly from the identification of a subset of relevant variables. Dimension reduction techniques, like Principal Component Analysis (PCA), projection to latent structures (PLS-DA), single class modeling (SIMCA) and kernel methods (SVM) are generally adopted to this aim. In some fields of application, like

in food-authentication, mislabeled and adulterated spectra may appear both in the calibration and/or validation sets. This contamination produces dramatic effects on the model estimation, and consequently on its prediction accuracy. To overcome this issue, a recent proposal in the literature introduces a variable selection step within the Robust Eigenvalue Decomposition Discriminant Analysis framework (Cappozzo *et al.*, 2019). Under the realistic assumption that only a portion of the spectral region is relevant for class discrimination, the procedure i) robustly identifies a subset of wavenumbers onto which building the decision rule, ii) protects it from potential label noise and outliers, and iii) simultaneously identifies anomalous samples.

We will recall here the main idea onto which the stepwise algorithm works, redirecting the interested reader to Cappozzo *et al.* (2021) for a more detailed presentation. The detection of p relevant features (out of the whole collection of $P \gg p$ available variables) on which to train the classifier has many advantages. Firstly, parameter estimation and interpretation is enhanced; secondly, loss on predictive power due to the inclusion of irrelevant and redundant information is avoided. Finally, cost reduction on future data collection and processing is obtained.

In model-based discriminant analysis, the features that directly depend on the class membership itself are called *relevant* variables. Conversely, *irrelevant* or noisy variables do not contain any discriminating power. Their distribution is completely independent on the group structure. Lastly, *redundant* variables essentially contain discriminant information that is already provided by the relevant ones: their distribution is conditionally independent of the grouping variable, given the relevant ones.

The algorithm starts from the empty set and, at each iteration, the inclusion of a *relevant* variable into the model is evaluated, based on its robustly assessed discriminating power. In a similar fashion, the removal of an existing variable from the model is also considered. The procedure iterates between variable addition and removal until two consecutive steps have been rejected.

2 Stability study

In this section, the results of a bootstrap-based analysis will be presented using data produced using non-parametric re-sampling of the actual data. The aim is to investigate the stability of the variable selection procedure.

The data we analyze come from the chemometric challenge organized during the "Chimiométrie 2005" conference (Fernández Pierna & Dardenne, 2007). The learning scenario encompasses $N = 215$ training and $M = 43$ test

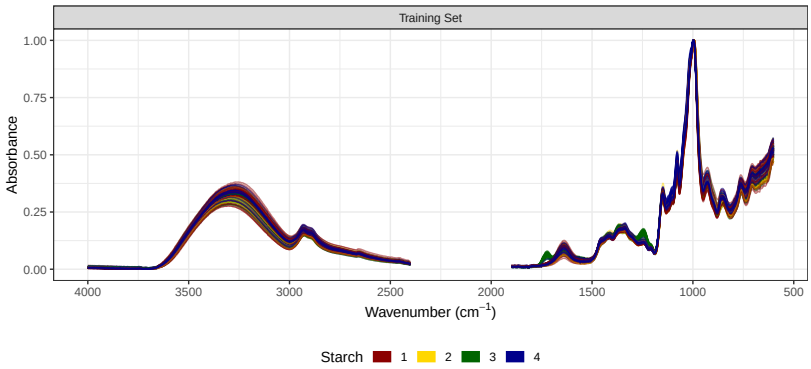

Figure 1. *Starches dataset: mid-infrared spectra of four starches classes.*

MIR spectra of starches of $G = 4$ different classes. For each sample, a total of $P = 2901$ absorbance measurements are recorded. A subset of training observations is displayed in Fig. 1. The aim of the competition was to discriminate the four different groups, defining a classification rule from the training set. In addition, outlier detection was advisable: four intentionally corrupted spectra were manually placed in the test set, as described in Fernández Pierna & Dardenne (2007).

For the first experiment 100 bootstrap datasets, of the same size as the actual dataset, were generated by sampling with replacement from the training set. A pattern in the selected variables arises from our results. For each bootstrapped sample, all models were fitted and the best-fit model was chosen using the BIC criterion, and the selected wavelengths were recorded. The chosen wavelengths show us which parts of the spectrum are of importance when classifying samples into different starches types. Results are shown in Fig. 2 through a raster plot. As we expect, there is some variability, due to the fact that the role of "relevant" and "irrelevant" variable is judged in terms of the set of already selected features. The wavelengths $997\,cm^{-1}$ and $995\,cm^{-1}$ correspond to spectral distributions of *amylose* and *amylopectin*, which are known to be present in different ratios across the starch classes. They have been selected with higher frequency, respectively 17 and 21 times in 67 runs.

3 Conclusions and further research

We developed a first stability analysis for a recent method for robust variable selection and classification, applied to spectrometric data. By a bootstrap simulation study on the learning set, although there has been variability in the

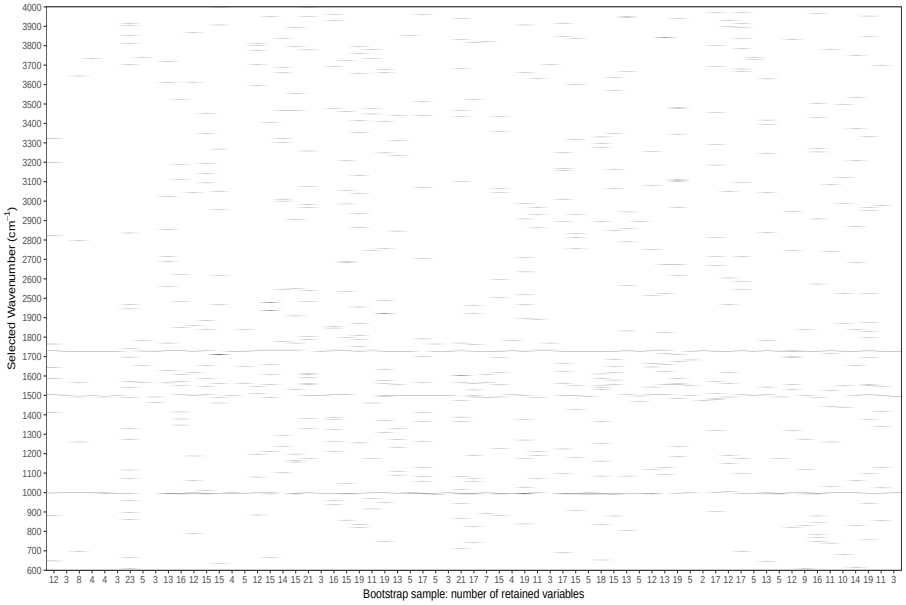

Figure 2. *Results of the stability analysis: for each of the 67 bootstrap samples, the selected wavenumbers are indicated in a raster plot.*

structure of the selected models, some stable pattern arises in results. Further research is still needed to cast more light on this topic. For instance, to investigate the sensitivity of the derived decision model, its accuracy on the test set is worth being analyzed, to establish the level of reliability in the resulting classification. This would mitigate the use of only a few real data examples and hence allows a more general discussion of the results.

References

CAPPOZZO, A., GRESELIN, F., & MURPHY, T. B. 2019. A robust approach to model-based classification based on trimming and constraints. *Advances in Data Analysis and Classification*, 1–28.

CAPPOZZO, A., DUPONCHEL, L., GRESELIN, F., & MURPHY, T. B. 2021. Robust variable selection in the framework of classification with label noise and outliers: Applications to spectroscopic data in agri-food. *Analytica Chimica Acta*, **1153**, 338245.

FERNÁNDEZ PIERNA, J. A., & DARDENNE, P. 2007. Chemometric contest at "Chimiométrie 2005": A discrimination study. *Chemometrics and Intelligent Laboratory Systems*, **86**(2), 219–223.