Robust variable selection in the framework of classification with label noise and outliers: applications to spectroscopic data in agri-food

Andrea Cappozzo* Ludovic Duponchel †
Francesca Greselin* Thomas Brendan Murphy ‡

Abstract

Classification of high-dimensional spectroscopic data is a common task in analytical chemistry. Well-established procedures like support vector machines (SVMs) and partial least squares discriminant analysis (PLS-DA) are the most common methods for tackling this supervised learning problem. Nonetheless, interpretation of these models remains sometimes difficult, and solutions based on wavelength selection are often preferred as they lead to clearer chemometrics interpretation. Unfortunately, for some delicate applications like food authenticity, mislabeled and adulterated spectra occur both in the calibration and/or validation sets, with dramatic effects on the model development, its prediction accuracy and robustness. Motivated by these issues, we propose to employ a robust model-based method for jointly performing variable selection and label noise detection. We demonstrate the effectiveness of our proposal in dealing with three agri-food spectroscopic studies, where several forms of perturbations are considered. Our approach succeeds in diminishing problem complexity, identifying anomalous spectra and attaining competitive predictive accuracy considering a very low number of selected wavelengths.

Keywords: Variable Selection; Robust classification; Label noise; Outlier detection; Near infrared spectroscopy; Mid infrared spectroscopy; Agri-food

1 Introduction

Near-infrared (NIR) and mid-infrared (MIR) spectroscopy have nowadays become a standard analytical practice in countless fields, being fast and non-invasive techiniques for promptly characterizing samples of interest [33, 41]. By acquiring a large number of absorbance values in a spectral range, NIR and MIR analyses provide compositional information for the products under study, with the final aim of being employed by chemometricians in developing multivariate

*Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy, andrea.cappozzo@unimib.it, francesca.greselin@unimib.it

†Univ. Lille, CNRS, UMR 8516 - LASIRE–Laboratoire avancé de spectroscopie pour les interactions, la réactivité et l'environnement, F-59000 Lille, France, ludovic.duponchel@univ-lille.fr

‡School of Mathematics & Statistics and Insight Research Centre, University College Dublin, Dublin, Ireland, brendan.murphy@ucd.ie

models. Generally, variables in the so-obtained feature space appear in the order of thousands, undermining the usage of standard low-dimensional techniques [44]. To this extent, variable selection methods play a pivotal role in determining a relevant subset of wavelengths onto which perform any subsequent analysis [6, 5]. Indeed, the detection of the most informative segments in a spectral region offer numerous advantages. Firstly, it reduces problem complexity, leading to faster and more interpretable models. Secondly, loss on predictive power is avoided by excluding the contribution of irrelevant and redundant noisy areas. Thirdly, cost impact for future data collection and processing will be reduced. Fourthly, robustness properties are conveyed into the estimation, for which the signal to noise ratio, as a by-product of the spectral selection, is automatically improved. Lastly, and most importantly, spectral interpretation is facilitated, whence chemometricians may uncover previously unknown properties and differences among the considered samples [21]. For all the aforementioned reasons, chemometrics literature has always been greatly benefited by variable selection methodologies, and recent examples include the successful determination of soil properties [42], yeast and oil concentration levels in beer and corn [45], yeast fermentation process using Raman spectroscopy [23], holocellulose and lignin content in multispecies hardwoods [27] and identification of adulterated Sanqi powder [12].

Conceptually, a variable selection method requires a) the definition of a relevance measure and b) the choice of an algorithm to perform the search. Standard procedures used in chemometrics, such as Competitive adaptive reweighted sampling [26], uninformative variable elimination [11], Monte Carlo-uninformative variable elimination [7], successive projections algorithm [1] and genetic algorithms [25] fall within this quite general paradigm. Despite the well-established effectiveness of the above-mentioned methods, all their data-dependent steps rely on the implicit assumption that samples are not affected by contamination. That is, the employed relevance measures are not robust against noisy observations, so much so that, when adulterations occur, the reliability of the entire output may be jeopardized. Thankfully, spectroscopic data are most often recorded in controlled experiments. Nevertheless, there exists some delicate applications, such as sample authenticity in agri-food, in which the raw material itself may be spoiled and/or adulterated [36]. Therefore, robust variable selection methods resistant to outliers and potential label noise are desirable. Particularly, the latter type of noise is seldom studied in analytical chemistry when developing a classification model, implicitly neglecting the circumstances in which such a situation may appear. Spectra with low interclass and high intraclass variability, inadequacy of low-cost automatic labeling systems and/or inexperienced personnel, label inconsistency when multiple experts are tasked to classify the same sample, information loss and data-entry errors are only some of the causes that are likely to lead to mislabeling.

Motivated by the preceding arguments, the present article illustrates the capabilities of a robust variable selection method, recently introduced in the literature [9], in performing high-dimensional classification in presence of label noise and outliers within a chemometrics context. Three successful applications to agri-food spectroscopic datasets will be analyzed and discussed. Specifically, the first part of the paper will briefly introduce the methodology and the datasets employed in the study. In the second part, model results will be presented, in comparison with state-of-the-art chemometric strategies. The

manuscript concludes with a discussion, highlighting how the advantages of the proposed method could positively impact classification of samples from spectroscopic data.

2 Material and methods

2.1 Robust variable selection: the stepwise REDDA approach

The methodology described in the present Section falls within the model-based family of classifiers, coupled with a novel variable selection procedure resistant to outliers and label noise. The main concepts underlying the method are hereafter reported.

Classification, also known as discriminant analysis, identifies the task of constructing a decision rule to assign an unlabeled sample to one of G known classes. For doing so, a complete set of N learning observations (i.e., the training set)

$$(\mathbf{x}, \mathbf{l}) = \left\{ (\mathbf{x}_1, \mathbf{l}_1), \ldots, (\mathbf{x}_N, \mathbf{l}_N); \mathbf{x}_n \in \mathbb{R}^P,\ \mathbf{l}_n = \{l_{n1}, \ldots, l_{nG}\}' \in \{0,1\}^G;\ n = 1, \ldots, N \right\} \tag{1}$$

is at our disposal; where $\mathbf{x}_n$ denotes a P-dimensional continuous predictor and $\mathbf{l}_n$ is its associated class label, such that $l_{ng} = 1$ if observation n belongs to group g and 0 otherwise with, clearly, $\sum_{g=1}^{G} l_{ng} = 1\ \forall n \in \{1, \ldots, N\}$. Specifically, in a spectroscopic dataset, P represents the total number of spectral variables in which the absorbance value is recorded for example. Model-based classifiers require some probabilistic assumptions in terms of the data-generating mechanism: we assume that the prior probability of class g is $\tau_g > 0$, $\sum_{g=1}^{G} \tau_g = 1$. The g-th class-conditional densities are independent P-dimensional Gaussian, with mean vector $\boldsymbol{\mu}_g \in \mathbb{R}^P$ and covariance matrix $\boldsymbol{\Sigma}_g \in PD(P)$: $\mathbf{x}_n | \mathbf{l}_{ng} = 1 \sim N_P(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$. The joint density of $(\mathbf{x}_n, \mathbf{l}_n)$ is therefore given by:

$$p(\mathbf{x}_n, \mathbf{l}_n; \boldsymbol{\theta}) = p(\mathbf{l}_n; \boldsymbol{\tau}) p(\mathbf{x}_n | \mathbf{l}_n; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = \prod_{g=1}^{G} \left[\tau_g \phi(\mathbf{x}_n; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right]^{l_{ng}} \tag{2}$$

where $\phi(\cdot; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ denotes the multivariate normal density and $\boldsymbol{\theta}$ represents the collection of parameters to be estimated, $\boldsymbol{\theta} = \{\tau_1, \ldots, \tau_G, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G\}$. Once the model has been fitted to the training set, test units $\mathbf{y}_m$, $m = 1, \ldots, M$, are assigned to the g-th class via the maximum a posteriori (MAP) rule:

$$\underset{g \in \{1, \ldots, G\}}{\arg\max} \frac{\hat{\tau}_g \phi\left(\mathbf{y}_m; \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g\right)}{\sum_{j=1}^{G} \hat{\tau}_j \phi\left(\mathbf{y}_m; \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j\right)}. \tag{3}$$

This formulation identifies a quite generic supervised classification device, and its effectiveness in defining decision rules for spectroscopic datasets has been reported in [43], [13], [40], [31] and [22], among others. For a general account on probabilistic model-based discriminant analysis and clustering methods in chemometrics, the reader is referred to the excellent review in [4].

Among the many specifications developed from the probabilistic structure in (2), the one considered here is the so called Eigenvalue Decomposition Discriminant Analysis (EDDA) [3]. EDDA defines a family of constrained models, where

different assumptions about the covariance matrices are imposed by considering the following eigenvalue decomposition:

$$\boldsymbol{\Sigma}_g = \lambda_g \boldsymbol{D}_g \boldsymbol{A}_g \boldsymbol{D}_g^{'} \tag{4}$$

where $\boldsymbol{D}_g$ is an orthogonal matrix of eigenvectors, determining groups orientation, $\boldsymbol{A}_g$ is a diagonal matrix such that $|\boldsymbol{A}_g| = 1$, accounting for groups shape, and $\lambda_g = |\boldsymbol{\Sigma}_g|^{1/p}$ is a scalar that controls the associated volume. By imposing some of the quantities in (4) to be equal across groups, the problem of over-parametrized modeling is mitigated. REDDA [8], a robust model-based classifier, was introduced to extend the EDDA framework to handle label noise and outliers. REDDA is based on the maximization of a *trimmed mixture log-likelihood* [32], where a trimming level γ assures that the most unlikely $\lfloor N\gamma \rfloor$ data points under the postulated model are discarded, ultimately robustifying parameter estimates. Nonetheless, despite the parsimonious structure induced by the eigen-decomposition in (4), for analytical spectroscopic applications the number of variables can be much greater than the number of observations, so much so that the REDDA model may still suffer from the curse of dimensionality [2], jeopardizing its performance in high-dimensional spaces. To overcome this limitation, a recent contribution in the literature proposes to include a variable selection step within the REDDA framework [9]: the core methodology employed in the present paper. Under the reasonable assumption that only a portion of the spectral region is relevant for class discrimination, the procedure robustly identifies a subset of wavenumbers onto which building a (robust) decision rule. The attained output is a method that performs high-dimensional classification with variable selection, safeguarding it from potential label noise and outliers, identifying such anomalous samples as a by-product of the learning process.

The devised stepwise algorithm works as follows: we start from the empty set and, at each iteration, the inclusion of an extra 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, then it stops. In details, at each iteration we partition the learning observations $\mathbf{x}_n$, $n = 1, \ldots, N$, into three parts $\mathbf{x}_n = (\mathbf{x}_n^c, x_n^p, \mathbf{x}_n^o)$, where:

- $\mathbf{x}_n^c$ indicates the set of variables currently included in the model
- x_n^p the variable proposed for inclusion
- $\mathbf{x}_n^o$ the remaining variables.

The intent here is to determine whether x_n^p shall be included (excluded) into (from) the relevant subset. To do so, we recast the problem as a model selection task, comparing the following two competing models:

- *Grouping ($\mathcal{M}_{GR}$):* $p(\mathbf{x}_n|\mathbf{l}_n) = p(\mathbf{x}_n^c, x_n^p, \mathbf{x}_n^o|\mathbf{l}_n) = p(\mathbf{x}_n^c, x_n^p|\mathbf{l}_n)p(\mathbf{x}_n^o|x_n^p, \mathbf{x}_n^c)$
- *No Grouping ($\mathcal{M}_{NG}$):* $p(\mathbf{x}_n|\mathbf{l}_n) = p(\mathbf{x}_n^c, x_n^p, \mathbf{x}_n^o|\mathbf{l}_n) = p(\mathbf{x}_n^c|\mathbf{l}_n)p(x_n^p|\mathbf{x}_n^r \subseteq \mathbf{x}_n^c)p(\mathbf{x}_n^o|x_n^p, \mathbf{x}_n^c)$

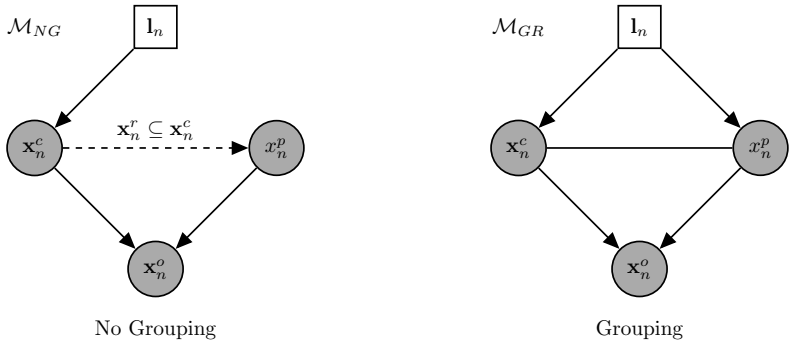

Figure 1: Graphical Representation of the *Grouping* and the *No Grouping* models

where $\mathbf{x}_n^r$ denotes a subset of the currently included variables $\mathbf{x}_n^c$. As ilustrated in Figure 1, $\mathcal{M}_{GR}$ assumes that x_n^p provides extra grouping information beyond that provided by $\mathbf{x}_n^c$; whereas $\mathcal{M}_{NG}$ specifies that x_n^p is conditionally independent of the group membership given $\mathbf{x}_n^r$. We consider $\mathbf{x}_n^r$ in the conditional distribution because x_n^p might be related to only a subset of the grouping variables $\mathbf{x}_n^c$ [29]. According to the general model-based structure described at the beginning of the Section, we assume $p(\mathbf{x}_n^c, x_n^p|\mathbf{l}_n)$ and $p(\mathbf{x}_n^c|\mathbf{l}_n)$ to be normal densities with constrained covariances, while $p(\mathbf{x}_n^o|x_n^p, \mathbf{x}_n^c)$ is only considered to be the same for both grouping and no grouping specification. Exploiting standard results for multivariate normal theory, (see, for example, Theorem 3.2.4 in [28]) $p(x_n^p|\mathbf{x}_n^r \subseteq \mathbf{x}_n^c)$ defines a normal linear regression model. More specifically, the involved models are of the form:

$$\mathcal{M}_{GR} : (\mathbf{x}_n^c, x_n^p), \mathbf{l}_n \sim \prod_{g=1}^{G} \left[\tau_g^{cp} \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_g^{cp}, \boldsymbol{\Sigma}_g^{cp})\right]^{l_{ng}}$$

$$\mathcal{M}_{NG} : \mathbf{x}_n^c, \mathbf{l}_n \sim \prod_{g=1}^{G} \left[\tau_g^{c} \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_g^{c}, \boldsymbol{\Sigma}_g^{c})\right]^{l_{ng}}, \quad x_n^p|\mathbf{x}_n^r \sim \mathcal{N}\left(\alpha + \boldsymbol{\beta}' \mathbf{x}_n^r, \sigma^2\right).$$

A standard way to perform model comparison is via the Bayes Factor ($\mathcal{B}_{GR,NG}$) [24], evaluating the plausibility of the *Grouping* model with respect to the *No Grouping* one. In the stepwise REDDA approach, a robust proxy to $\mathcal{B}_{GR,NG}$ is employed to select which specification to prefer. Following [35], twice the logarithm of $\mathcal{B}_{GR,NG}$ is approximated with

$$2\log\left(\mathcal{B}_{GR,NG}\right) \approx TBIC(GR) - TBIC(NG) \tag{5}$$

where the trimmed BIC (TBIC), firstly introduced in [32], acts as a robust version of the Bayesian Information Criterion [38] employed in the approximation in (5). Particularly, for the *Grouping* and the *No Grouping* specification outlined above, the TBICs respectively read:

$$TBIC(GR) = \underbrace{2\sum_{n=1}^{N}\zeta(\mathbf{x}_n^c, x_n^p)\sum_{g=1}^{G} l_{ng}\log\left(\hat{\tau}_g^{cp}\phi(\mathbf{x}_n^c, x_n^p; \hat{\boldsymbol{\mu}}_g^{cp}, \hat{\boldsymbol{\Sigma}}_g^{cp})\right)}_{2\times\text{trimmed log maximized likelihood of } p(\mathbf{x}_n^c, x_n^p, \mathbf{l}_n)} + \\ - v^{cp} log(N^*) \tag{6}$$

$$TBIC(NG) = \underbrace{2\sum_{n=1}^{N}\iota(\mathbf{x}_n^c, x_n^p)\sum_{g=1}^{G} l_{ng}\log\left(\hat{\tau}_g^{c}\phi(\mathbf{x}_n^c; \hat{\boldsymbol{\mu}}_g^{c}, \hat{\boldsymbol{\Sigma}}_g^{c})\right)}_{2\times\text{trimmed log maximized likelihood of } p(\mathbf{x}_n^c, \mathbf{l}_n)} - v^c log(N^*) + \\ \underbrace{+2\sum_{n=1}^{N}\iota(\mathbf{x}_n^c, x_n^p)\log\left[\phi\left(x_n^p; \hat{\alpha} + \hat{\boldsymbol{\beta}}^{'}\mathbf{x}_n^r, \hat{\sigma}^2\right)\right]}_{2\times\text{trimmed log maximized likelihood of } p(x_n^p | \mathbf{x}_n^r \subseteq \mathbf{x}_n^c)} - v^p log(N^*). \tag{7}$$

The quantities v^{cp} and v^c are penalty terms, namely the number of parameters for a REDDA model estimated on the set of variables $\mathbf{x}_n^c, x_n^p$ and $\mathbf{x}_n^c$, respectively; while v^p accounts for the number of parameters in the linear regression of x_n^p on $\mathbf{x}_n^r$. The terms $\zeta(\cdot)$ and $\iota(\cdot)$ are 0-1 indicator functions, identifying the subset of observations that have null weight in the trimmed likelihood under $\mathcal{M}_{GR}$ and $\mathcal{M}_{NG}$, with $N^* = \sum_{n=1}^{N}\zeta(\mathbf{x}_n) = \sum_{n=1}^{N}\iota(\mathbf{x}_n)$. That is, potential outlying and mislabeled observations do not influence the selection procedure, since only $N^* = \lceil N(1-\gamma)\rceil$ samples are accounted for parameters estimation, with γ denoting the impartial trimming level. The set of parameters $\{\alpha, \boldsymbol{\beta}, \sigma^2\}$ are related to the linear regression component, and are robustly estimated via maximum likelihood on the untrimmed samples.

In the addition stage, the x_n^p variable with highest positive difference in (5) (if any) is the one selected for inclusion. In the removal stage, x_n^p takes the role of the variable to be dropped, and the one displaying highest positive difference in (5) (if any) is excluded from the set of currently included variables $\mathbf{x}_n^c$. When neither addition nor removal move is performed, the procedure terminates. In this way, the number of relevant variables necessary to build the classification rule is automatically inferred, and it needs not be a priori specified. The routines for the stepwise REDDA approach have been written in `R` language [34]: the source code is openly available at `https://github.com/AndreaCappozzo/varselTBIC`.

2.2 Spectroscopic datasets in agri-food

The stepwise REDDA approach is applied to the analysis of three different multiclass data sets. The first one is a 4-class problem where the observations are mid-infrared spectra of modified starches. The second one is a 5-class problem encompassing visible and near infrared reflectance spectra of homogenized meat samples. The last dataset is a 2-class problem, concerning the discrimination between Ligurian and Non-Ligurian olive oil through mid-infrared measurements. For the considered datasets, employed instrumentation and sample collection procedures are thoroughly described in [18], [30] and [20]; thus, only a succinct explanation will be hereafter reported. All these challenging situations represent

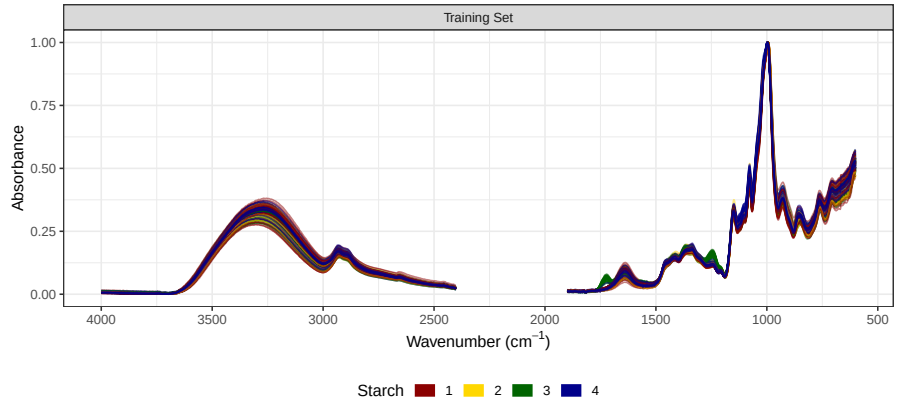

Figure 2: Mid-infrared spectra of four starches classes. Starches dataset.

typical high-dimensional classification tasks often encountered in spectroscopy: a variable selection procedure resistant to label noise and outliers can potentially be beneficial in this regard. As a last worthy note, we mention that the subsequent analyses are directly performed on the raw spectra, without any pretreatment applied to the samples originally provided.

Starches data set

The first dataset comes from the chemometric challenge organized during the 'Chimiométrie 2005' conference [17]. The learning scenario encompasses $N = 215$ training and $M = 43$ test 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 Figure 2. The participants of the competition were tasked to discriminate as accurately as possible the four different classes, defining a classification rule from the training set. In addition, outlier detection needed to be performed, as four intentionally corrupted spectra were manually placed in the test set: a graphical representation is depicted in Figure 3. For a thorough description on how these modifications were obtained, the interested reader is referred to [17]. In addition, we slightly complicate the learning framework even further including less than 2% of label noise in the training set: the last four samples of the third class are wrongly labeled as coming from the fourth one.

Meat data set

The second dataset reports the NIR spectra of 231 homogenized meat samples, recorded from $400 - 2498$ nm at intervals of 2 nm, accounting for a total of $P = 1050$ spectral variables. Spectra belong to five different meat types, with 32 beef, 55 chicken, 34 lamb, 55 pork, and 55 turkey. We randomly partition the recorded spectra into calibration and test sets: the former is composed by 16 beef, 28 chicken, 17 lamb, 28 pork and 28 turkey (the resulting training set is displayed in Figure 4), while the latter contains the same proportion of these five meat types with four additional spectra manually adulterated as follows:

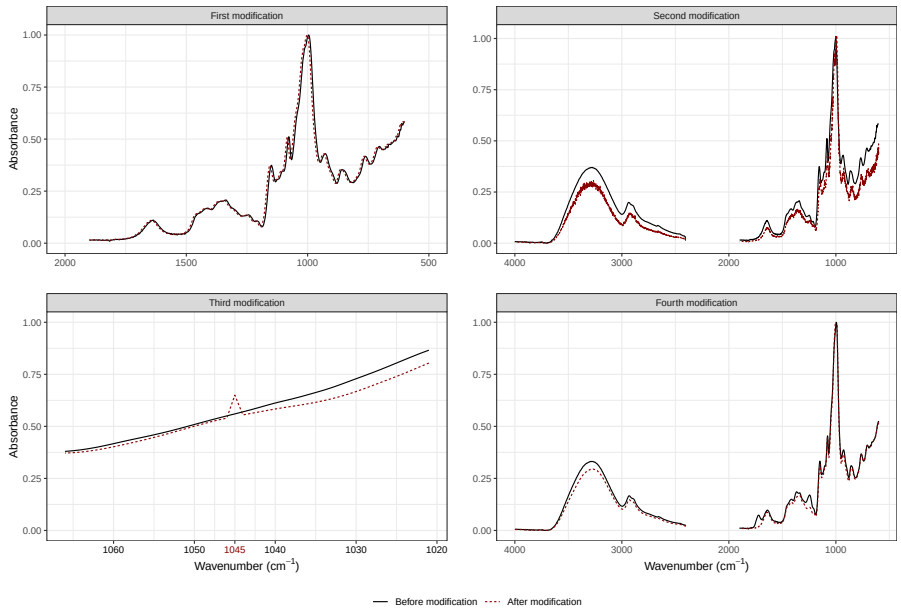

Figure 3: The 4 adulterated spectra manually placed in the test set by the 'Chimiométrie 2005' contest organizers, before and after modification. Starches dataset.

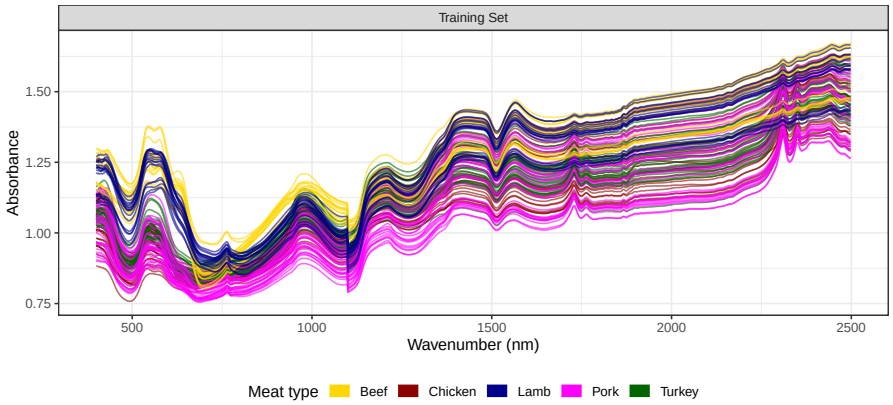

Figure 4: Visible and near infrared spectra of five homogenized meat types. Meat dataset.

- a shifted version of a pork spectrum, achieved by removing the first 15 data points and appending the last 15 group-mean absorbance values at the end of it;
- a noisy version of a pork spectrum, generated by adding Gaussian white noise to the original one;
- a modified version of a turkey spectrum, obtained by abnormally increasing the absorbance value in a single specific wavelength to simulate a spike;
- a pork spectrum with an added slope, produced by multiplying the original spectrum by a positive constant.

These modifications mimic the ones considered in the "Chimiométrie 2005" chemometric contest for the starches dataset, described in the previous Section, and agree with those reported in [14] within a novelty detection framework.

Olive Oil data set

The last dataset examines MIR olive oil spectra based on Fourier-transform (FTIR) measurements, with a learning scenario encompassing training and test sets of sizes respectively equal to $N = 280$ and $M = 630$. The aim here is to identify whether or not samples originate from the Italian coastal region of Liguria. In doing so, two nested subsets of wavelengths are considered in the analysis: the first one comprises the spectral zones from 3000 to 2400 cm^{-1} and from 2250 to 700 cm^{-1} ($P = 1117$ recorded features), while the second one covers the entire $4000 - 700$ cm^{-1} spectral range ($P = 1712$). The two resulting training sets are respectively displayed in the top and bottom panels of Figure 5, where red lines denote Ligurian olive oil spectra. Specifically, only the former subset was previously studied [20, 15]; since the absorption of atmospheric carbon dioxide was registered in the frequency range $2400 - 2250$ cm^{-1}, whereas the end of the spectrum seemed to contain mainly noise and was thus removed too. The reason for confronting with both scenarios is twofold. On the one hand, we aim at evaluating how the presence of additional noisy variables, in an already high dimensional problem, impacts the performance of our and competing methods. On the other hand, we are interested in assessing whether a knowledge-based selection, a common practice in chemometrics studies [37, 10, 46], is still unavoidable even when algorithmic procedures could take over such manual approach.

3 Results and discussion

Most classification problems in chemometrics cannot be solved by directly applying model-based classifiers, since $N \ll P$: the agri-food applications considered in this paper make no exception. To overcome this issue, we make use of the stepwise REDDA method previously introduced to provide a natural solution for dealing with contaminated high-dimensional data, and, as we will see, to identify adulterated spectra in the different scenarios.

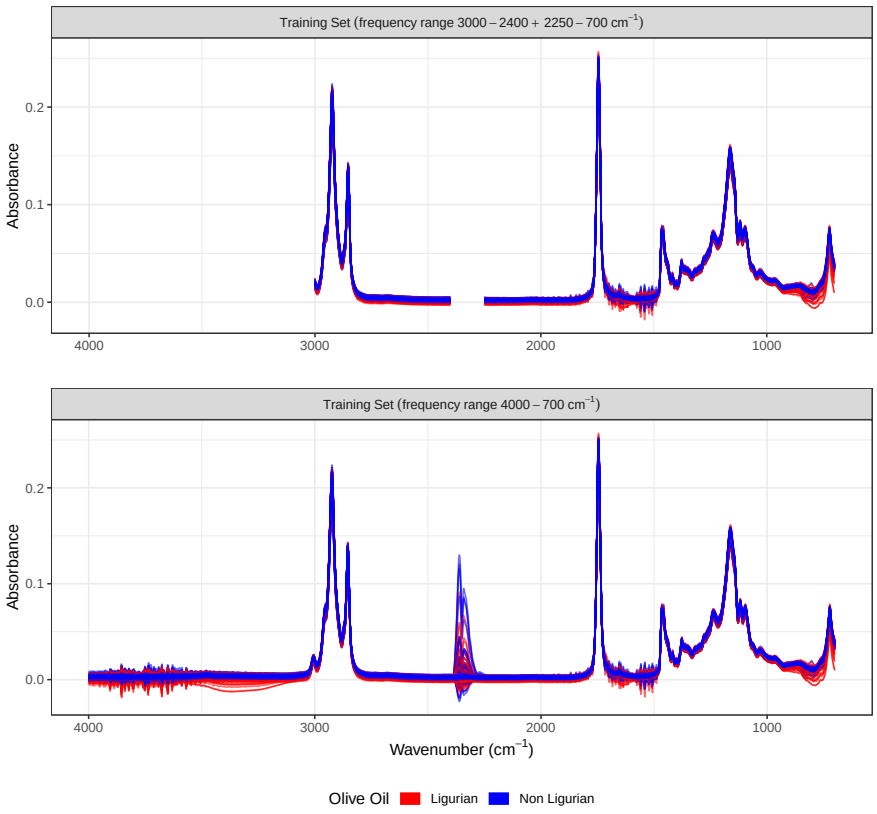

Figure 5: Mid-infrared spectra of Ligurian and Non Ligurian olive oil, spectral zones $3000 - 2400$ cm^{-1} + $2250 - 700$ cm^{-1} (top panel) and $4000 - 700$ cm^{-1} (bottom panel). Olive Oil dataset.

Starches dataset

For the starches dataset, we run the stepwise REDDA with $\gamma = 0.03$. That is, the method is protected against potential contaminated spectra in the training set, as only $100\lceil 1-\gamma \rceil\%$ of the samples is employed for model fitting, leaving the least likely $100\gamma\%$ unmodeled. Such robustification effectively takes care of the label noise placed in the calibration set, preventing it to spoil the wavelength selection. The procedure, out of $P = 2901$, selects a total of only six relevant spectral variables: 1728 cm^{-1}, 1682 cm^{-1}, 1555 cm^{-1}, 1502 cm^{-1}, 997 cm^{-1} and 995 cm^{-1}. The last two wavenumbers correspond to spectral distributions of amylose and amilopectin, which are known to be present in different ratios for the different starch classes. The other wavenumbers on the list correspond to very low levels of absorbance in the spectra which makes molecular interpretation difficult. Figure 6 displays the generalized pairs plot [16] for the selected variables. Such graphical tool encompasses different plot types depending on the paired combinations of categorical and/or quantitative variables, generalizing the standard scatterplot, depicted only in the lower triangular matrix. Graphs above the main diagonal report contours of 2D density estimates, where it stands out that the most difficult task resides in separating starch classes 1 and 2, as was already pointed out in [17]. Right and bottom margins respectively include side-by-side boxplots and faceted-density plots, useful in revealing patterns when dealing with one categorical and one continuous variable. Lastly, univariate plots are displayed in the main diagonal, namely density plots and a bar chart illustrating samples proportion.

A REDDA model with $\gamma = 0.03$ is employed to classify the test samples, using as predictors the spectral frequencies retained by the stepwise variable selector. A Support Vector Machine with Gaussian radial kernel (SVM) was also considered, as it was shown to be the best performing classifier for this specific dataset [18, 17]. In addition, we replicate the second best solution proposed by one of the 'Chimiométrie 2005' contest participants: an ensemble method was constructed by combining ROC, PLS and SVM predictions via majority vote on a subset of variables, previously determined by a PLS model. Classification accuracy for the three competing methods, learned on both the original training set and on the one containing label noise, are reported in Table 1. Our

Table 1: Number of correctly predicted test samples and associated misclassification error for different methods, starches dataset. The test set without outliers has a total sample size of $M = 39$. Results with superscript * were originally reported in [17].

	Stepwise REDDA	SVM radial kernel	ROC+PLS+SVM
Training set with label noise			
# correctly predicted	32	31	31
% correctly predicted	82.1	79.5	79.5
Training set without label noise			
# correctly predicted	32	37*	33
% correctly predicted	82.1	94.9*	84.6

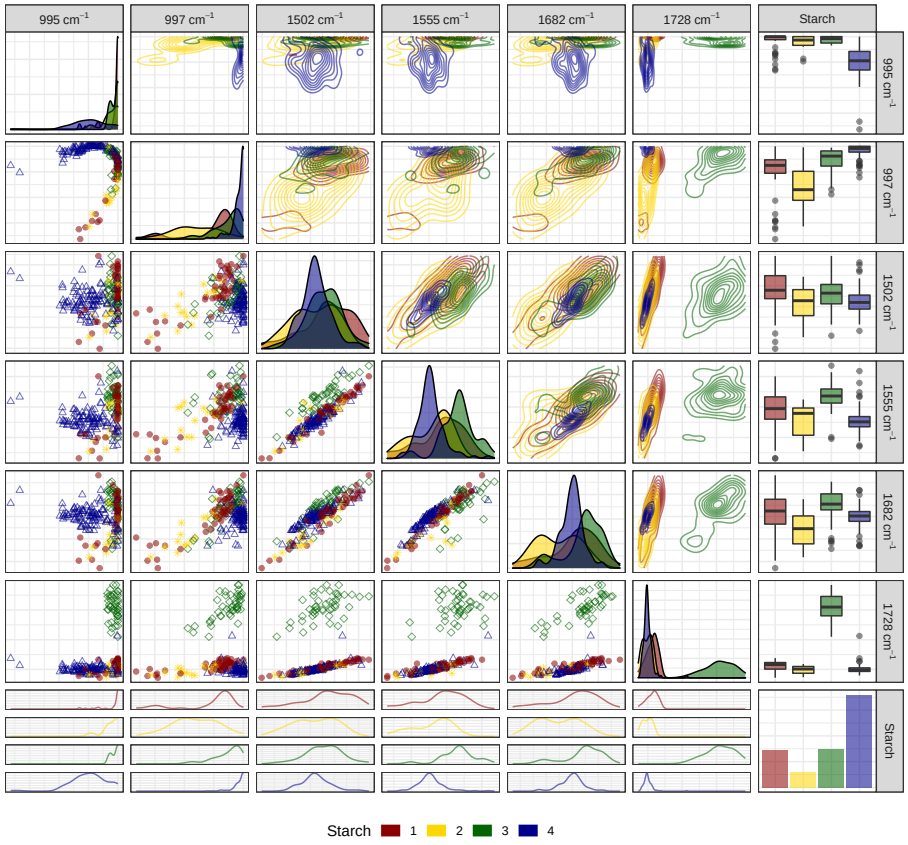

Figure 6: Generalized pairs plot of the spectral frequencies selected by the stepwise REDDA method. Starches dataset, training set.

robust model-based classifier attains the same predictive power when trained on either dataset: this is due to the γ level of trimming, that correctly identifies the adulterated spectra as to be label noise and thus safeguarding parameters from estimation bias. On the other hand, the performance of the kernel and ensemble methods are negatively impacted by the presence of the 4 mislabeled samples. Our proposal maintains intact predictive power, further providing a great reduction in model complexity and results interpretation, being the final procedure only based on $p = 6$ wavenumbers. The tremendous decrease in data dimension, together with the ability of successfully dealing and identifying label noise are two most desirable aspects in chemometrics. Overall, the selection of only six frequencies seems sufficient to well-capture the heterogeneity in the starches population.

We mentioned at the beginning of the previous Section that 4 adulterated spectra were manually placed in the test set (see Figure 3). While the performance of the different methods has been evaluated on the clean units only to assure fairness in the comparison, our methodology can be further employed to perform outlier detection considering the estimated marginal density for each test unit $\mathbf{y}_m$:

$$\hat{p}(\mathbf{y}_{m,\hat{F}}; \hat{\tau}, \hat{\boldsymbol{\mu}}_{\hat{F}}, \hat{\boldsymbol{\Sigma}}_{\hat{F}}) = \sum_{g=1}^{G} \hat{\tau}_g \phi \left(\mathbf{y}_{m,\hat{F}}; \hat{\boldsymbol{\mu}}_{g,\hat{F}}, \hat{\boldsymbol{\Sigma}}_{g,\hat{F}} \right) \tag{8}$$

where $\hat{F}$ denotes the relevant variables identified by the stepwise REDDA approach. The 3 spectra with lowest value of (8) are actually outliers. The only neglected anomaly is the one that was contaminated with a spike on a single wavelength, not identified as relevant by the feature selection method. Consequently, its marginal density in (8) is not altered by the manual modification. All things considered, our approach is able to effectively identify 3 out of 4 outliers and to greatly decrease problem complexity, whilst still maintain competitive predictive power when compared with state-of-the-art classifiers.

Meat dataset

The stepwise REDDA procedure is applied to the meat dataset: the aim here is to discriminate the five meat types as well as to identify the 4 anomalous spectra, manually placed in the validation set. The obtained classification on the genuine test samples (without considering the 4 adulterated ones) is reported in Table 2, achieving a misclassification error of 6.14%. A remarkably good performance is exhibited by our proposal, in agreement with results obtained by most advanced methods, whose performances are reported in Table 2: the interested reader is referred to [31, 19, 39] for the associated classification studies. Our methodology, out of $P = 1050$, selects a total of six relevant wavelengths: 636 nm, 704 nm, 870 nm, 1076 nm, 668 nm and 674 nm. Such wavelengths span a spectral region related to proteins. A generalized pairs plot is reported in Figure 7, where we observe how the spectral variables selected by our method indeed reveal distinctive patterns among the meat types, even though the poultry classes, namely chicken and turkey, are still difficult to distinguish one another. Similarly to what done in the previous subsection for the starches dataset, the marginal density defined in (8) can be used to assess the presence of outlying units in the test set: the 4 samples with lowest value of (8)

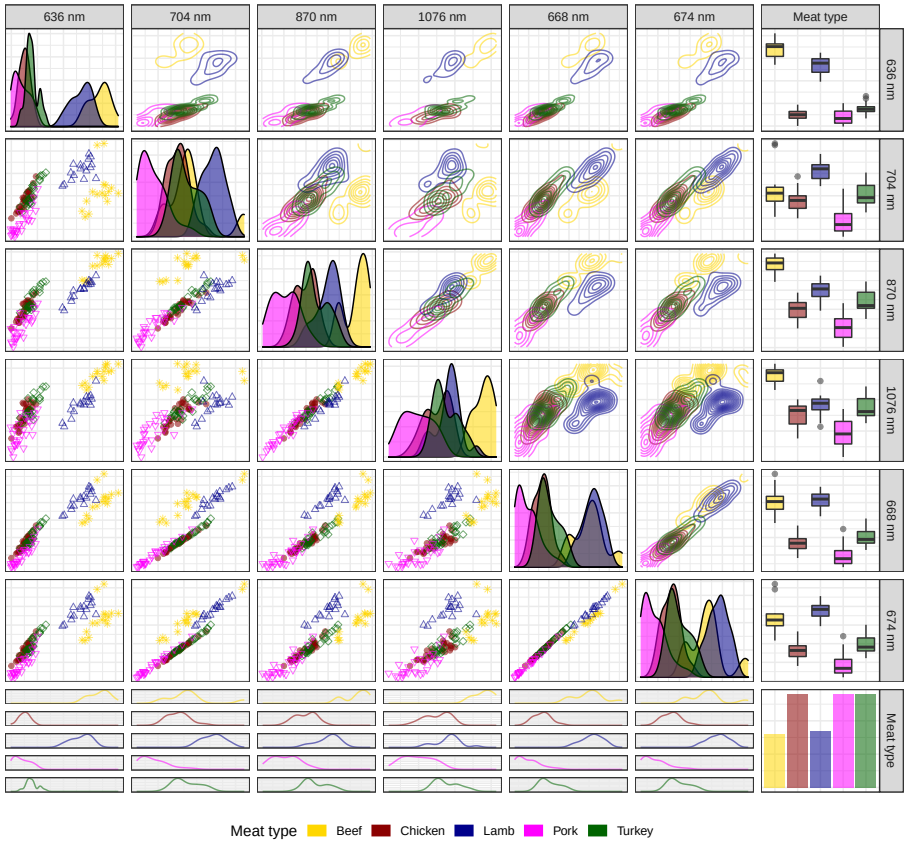

Figure 7: Generalized pairs plot of the spectral frequencies selected by the stepwise REDDA method. Meat dataset, training set.

Table 2: Number of correctly predicted test samples and associated classification accuracy for different methods, meat dataset. Results with superscript *, superscript † and superscript ‡ were respectively reported in [31], [19] and [39].

	Stepwise REDDA	Murphy et al [31]	Gutierrez et al [19]	PLS-DA [39]
# correctly predicted	107	107*	-	-
% correctly predicted	93.9	93.9*	87.4†	94‡

are precisely the manually adulterated ones. For the meat dataset, exceeding the already good results shown in the starches analysis, outliers detection is thoroughly accomplished by means of the proposed approach.

Olive Oil dataset

Table 3: Number of correctly predicted test samples and associated classification accuracy for different methods on the two subsets of wavelengths, olive oil dataset. The test set has a total sample size of $M = 630$.

	Stepwise REDDA	SVM radial kernel	PLS-DA
Frequencies $3000 - 2400 + 2250 - 700$ cm^{-1}			
# correctly predicted	507	459	509
% correctly predicted	80.5	72.9	80.8
Frequencies $4000 - 700$ cm^{-1}			
# correctly predicted	505	428	503
% correctly predicted	80.2	67.9	79.8

Two distinct analyses, depending whether the reduced or the full spectral range (see Figure 5) is employed for model fitting, are accomplished for the olive oil dataset. Classification accuracy for both scenarios is reported in Table 3, for which stepwise REDDA, partial least squares discriminant analysis (PLS-DA) and SVM classifiers have been considered. Results displayed in the table highlight some peculiarities that are worth examining. In the first place, PLS-DA and SVM are negatively impacted by the roughly 600 more features in the full spectra case, where particularly the kernel method shows a considerable reduction in terms of predictive power. Contrarily, stepwise REDDA does not seem to be affected by the original size of the feature space, showcasing essentially unchanged classification accuracy for both spectra ranges. With reference to it, the wavenumbers selected by the procedure amount to frequencies 704 cm^{-1}, 1279 cm^{-1} and 1726 cm^{-1} when a-priori knowledge-based selection is accomplished; and to 1447 cm^{-1}, 1726 cm^{-1}, 3366 cm^{-1}, 3576 cm^{-1} and 3996 cm^{-1} in the full range scenario. In the first experiment, the relevant wavelengths correspond respectively to the $C - H$ bending of the group, to $C - C$ and $C - O$ bending situations, and to the streching of the carbonyl groups; while 3366 cm^{-1}, 3576 cm^{-1} in the second study are associated with $O-H$ bond contributions. In both situations, there is a consistent reduction in terms of problem dimension, correspondingly moving from 1117 and 1712 to 3 and 5 retained features. Figures

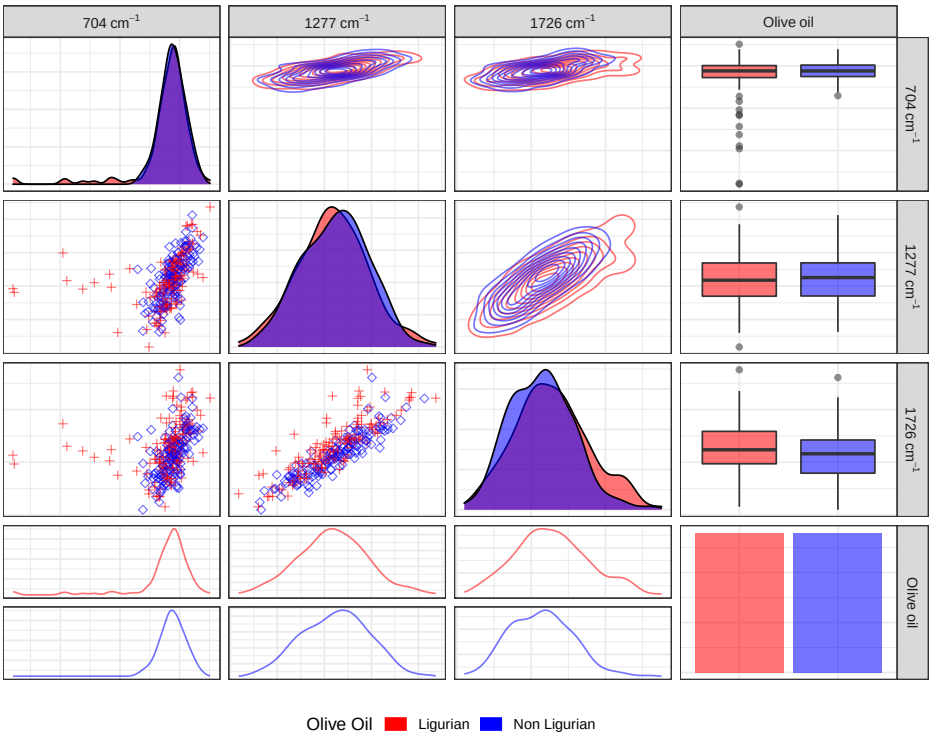

Figure 8: Generalized pairs plot of the spectral frequencies selected by the stepwise REDDA method. Reduced olive oil dataset (frequencies $3000 - 2400 + 2250 - 700$ cm^{-1}), training set.

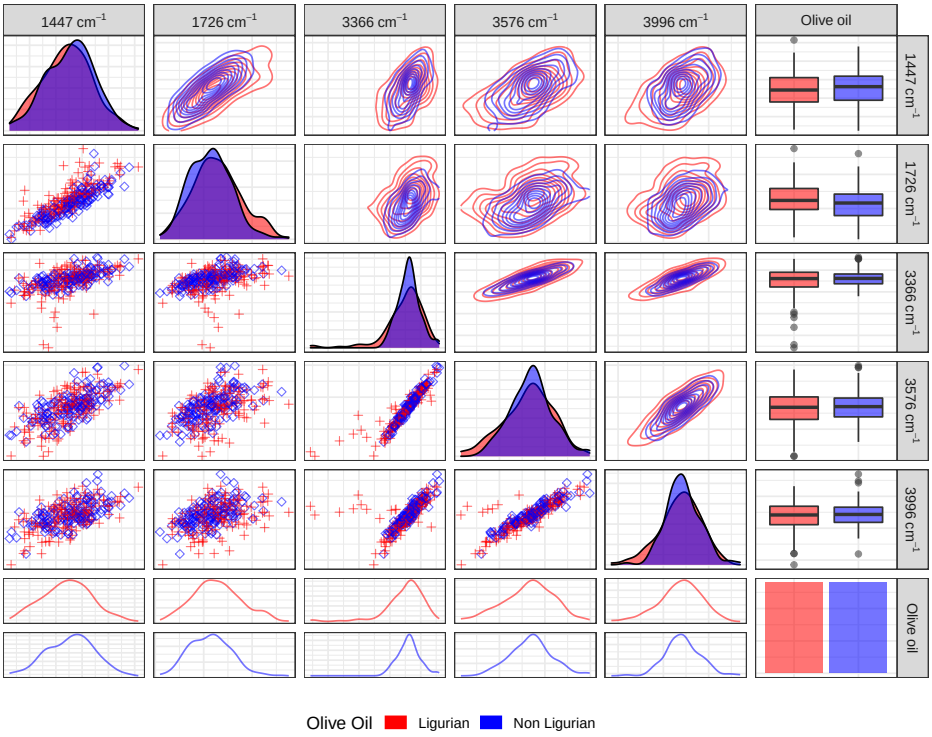

Figure 9: Generalized pairs plot of the spectral frequencies selected by the stepwise REDDA method. Full olive oil dataset (frequencies $4000 - 700$ cm^{-1}), training set.

8 and 9 display the pairs plots associated to such subsets of relevant variables, in which it is apparent that this discrimination task is harder than the previous ones, with classes separation much less discernible. While, as expected, no wavelength was selected in the range $2400 - 2250$, known to be contaminated by atmospheric carbon dioxide, three (3366 cm^{-1}, 3576 cm^{-1} and 3996 cm^{-1}) out of the five variables deemed to be relevant in the full scenario were manually discarded while defining the knowledge-based reduced dataset [20]. This unexpected result should not come as a surprise: indeed, by inspecting both the bottom panel in Figure 5 and the pairs plot in Figure 9, it is evident that wavelengths greater than 3000 cm^{-1} do possess some discriminating power. Thereupon, we argue that manual based spectral range election shall be performed with care, as some valuable information may be inadvertently lost.

4 Conclusion

The aim of the paper has been to showcase the benefits of a robust variable selection method for classification in chemometrics. Specifically, motivated by three agri-food applications, we have investigated the effect that contamination produces in standard tools for spectroscopic analysis, and how the proposed methodology can cope with it. Identifying noise as a factor that makes class dis-

crimination more challenging, we have confronted label noise (starches dataset), attribute noise (meat dataset) and noisy variables (olive oil dataset). Excellent results have been obtained in all three scenarios, wherein our robust feature selection has attained a reduction in problem complexity and accurate detection of mislabeled and/or adulterated spectra, whilst maintaining competitive predictive power. In addition, we have demonstrated that our method is directly applicable to raw spectra, without needing any preprocessing step.

Mislabeling is an issue oftentimes overlooked in analytical chemistry: a method that accomplishes variable selection, eliminating the need of manual approaches, while automatically protecting against potential contamination seems particularly desirable. Furthermore, the uncovering of the most discriminative frequencies or wavenumbers both facilitates chemometrics interpretation and generates drastic cost reduction. As a consequence, laser diodes can be employed to record only targeted wavenumbers, without the need to acquire, process and store the whole spectra.

In conclusion, based on our findings, we believe that the proposed procedure could be well-accepted by the chemometric community, while additional analyses may further validate its applicability in the spectroscopic field.

Acknowledgments

Brendan Murphy's work is supported by Science Foundation Ireland grants (SFI/12/RC/2289_P2 and 16/RC/3835)

References

[1] M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvão, T. Yoneyama, H. C. Chame, and V. Visani. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems*, 57(2):65–73, 2001.

[2] R. Bellman. *Dynamic Programming*. Rand Corporation research study. Princeton University Press, 1957.

[3] H. Bensmail and G. Celeux. Regularized Gaussian Discriminant Analysis through Eigenvalue Decomposition. *Journal of the American Statistical Association*, 91(436):1743–1748, dec 1996.

[4] C. Bouveyron. Probabilistic model-based discriminant analysis and clustering methods in chemometrics. *Journal of Chemometrics*, 27(12):433–446, dec 2013.

[5] J. M. Brenchley, U. Hörchner, and J. H. Kalivas. Wavelength Selection Characterization for NIR Spectra. *Applied Spectroscopy*, 51(5):689–699, may 1997.

[6] P. J. Brown. Wavelength selection in multicomponent near-infrared calibration. *Journal of Chemometrics*, 6(3):151–161, may 1992.

[7] W. Cai, Y. Li, and X. Shao. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 90(2):188–194, feb 2008.

[8] A. Cappozzo, F. Greselin, and T. B. Murphy. A robust approach to model-based classification based on trimming and constraints. *Advances in Data Analysis and Classification*, 14(2):327–354, jun 2020.

[9] A. Cappozzo, F. Greselin, and T. B. Murphy. Robust variable selection for model-based learning in presence of adulteration. jul 2020.

[10] M. Casale, M.-J. Sáiz Abajo, J.-M. González Sáiz, C. Pizarro, and M. Forina. Study of the aging and oxidation processes of vinegar samples from different origins during storage by near-infrared spectroscopy. *Analytica Chimica Acta*, 557(1-2):360–366, jan 2006.

[11] V. Centner, D.-L. Massart, O. E. de Noord, S. de Jong, B. M. Vandeginste, and C. Sterna. Elimination of Uninformative Variables for Multivariate Calibration. *Analytical Chemistry*, 68(21):3851–3858, jan 1996.

[12] H. Chen, C. Tan, and H. Li. Untargeted identification of adulterated Sanqi powder by near-infrared spectroscopy and one-class model. *Journal of Food Composition and Analysis*, 88, 2020.

[13] N. Dean, T. B. Murphy, and G. Downey. Using unlabelled data to update classification rules with applications in food authenticity studies. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 55(1):1–14, 2006.

[14] F. Denti, A. Cappozzo, and F. Greselin. A Two-Stage Bayesian Nonparametric Model for Novelty Detection with Robust Prior Information. 2020.

[15] O. Devos, G. Downey, and L. Duponchel. Simultaneous data pre-processing and SVM classification model selection based on a parallel genetic algorithm applied to spectroscopic data of olive oils. *Food Chemistry*, 148:124–130, 2014.

[16] J. W. Emerson, W. A. Green, B. Schloerke, J. Crowley, D. Cook, H. Hofmann, and H. Wickham. The generalized pairs plot. *Journal of Computational and Graphical Statistics*, 22(1):79–91, 2013.

[17] J. A. Fernández Pierna and P. Dardenne. Chemometric contest at 'Chimiométrie 2005': A discrimination study. *Chemometrics and Intelligent Laboratory Systems*, 86(2):219–223, apr 2007.

[18] J. A. Fernández Pierna, P. Volery, R. Besson, V. Baeten, and P. Dardenne. Classification of Modified Starches by Fourier Transform Infrared Spectroscopy Using Support Vector Machines. *Journal of Agricultural and Food Chemistry*, 53(17):6581–6585, aug 2005.

[19] L. Gutiérrez, E. Gutiérrez-Peña, and R. H. Mena. Bayesian nonparametric classification for spectroscopy data. *Computational Statistics and Data Analysis*, 78:56–68, 2014.

[20] S. Hennessy, G. Downey, and C. P. O' Donnell. Confirmation of Food Origin Claims by Fourier Transform Infrared Spectroscopy and Chemometrics: Extra Virgin Olive Oil from Liguria. *Journal of Agricultural and Food Chemistry*, 57(5):1735–1741, mar 2009.

[21] U. Indahl and T. Næs. A variable selection strategy for supervised classification with continuous spectroscopic data. *Journal of Chemometrics*, 18(2):53–61, feb 2004.

[22] J. Jacques, C. Bouveyron, S. Girard, O. Devos, L. Duponchel, and C. Ruckebusch. Gaussian mixture models for the classification of high-dimensional vibrational spectroscopy data. *Journal of Chemometrics*, 24(11-12):719–727, nov 2010.

[23] H. Jiang, W. Xu, Y. Ding, and Q. Chen. Quantitative analysis of yeast fermentation process using Raman spectroscopy: Comparison of CARS and VCPA for variable selection. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, 228, 2020.

[24] R. E. Kass and A. E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773, jun 1995.

[25] R. Leardi, R. Boggia, and M. Terrile. Genetic algorithms as a strategy for feature selection. *Journal of Chemometrics*, 6(5):267–281, sep 1992.

[26] H. Li, Y. Liang, Q. Xu, and D. Cao. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Analytica Chimica Acta*, 648(1):77–84, aug 2009.

[27] L. Liang, L. Wei, G. Fang, F. Xu, Y. Deng, K. Shen, Q. Tian, T. Wu, and B. Zhu. Prediction of holocellulose and lignin content of pulp wood feedstock using near infrared spectroscopy and variable selection. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, 225, 2020.

[28] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate analysis*. Academic Press London; New York, 1979.

[29] C. Maugis, G. Celeux, and M. L. Martin-Magniette. Variable selection in model-based discriminant analysis. *Journal of Multivariate Analysis*, 102(10):1374–1387, 2011.

[30] J. McElhinney, G. Downey, and T. Fearn. Chemometric Processing of Visible and near Infrared Reflectance Spectra for Species Identification in Selected Raw Homogenised Meats. *Journal of Near Infrared Spectroscopy*, 7(3):145–154, jun 1999.

[31] T. B. Murphy, N. Dean, and A. E. Raftery. Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. *The Annals of Applied Statistics*, 4(1):396–421, mar 2010.

[32] N. Neykov, P. Filzmoser, R. Dimova, and P. Neytchev. Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics and Data Analysis*, 52(1):299–308, sep 2007.

[33] C. Pasquini. Near infrared spectroscopy: A mature analytical technique with new perspectives – A review, 2018.

[34] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[35] A. E. Raftery and N. Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178, 2006.

[36] L. M. Reid, C. P. O'Donnell, and G. Downey. Recent technological advances for the determination of food authenticity. *Trends in Food Science & Technology*, 17(7):344–353, jul 2006.

[37] H. Sato, M. Kiguchi, F. Kawaguchi, and A. Maki. Practicality of wavelength selection to improve signal-to-noise ratio in near-infrared spectroscopy. *NeuroImage*, 21(4):1554–1562, apr 2004.

[38] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, mar 1978.

[39] M. Singh and K. Domijan. Comparison of Machine Learning Models in Food Authentication Studies. In *2019 30th Irish Signals and Systems Conference (ISSC)*, pages 1–6. IEEE, jun 2019.

[40] D. Toher, G. Downey, and T. B. Murphy. A comparison of model-based and regression classification techniques applied to near infrared spectroscopic data in food authentication studies. *Chemometrics and Intelligent Laboratory Systems*, 89(2):102–115, nov 2007.

[41] R. Valand, S. Tanna, G. Lawson, and L. Bengtström. A review of Fourier Transform Infrared (FTIR) spectroscopy used in food adulteration and authenticity investigations, 2020.

[42] M. Vohland, M. Ludwig, S. Thiele-Bruhn, and B. Ludwig. Determination of soil properties with visible to near- and mid-infrared spectroscopy: Effects of spectral variable selection. *Geoderma*, 223-225(1):88–96, jul 2014.

[43] W. Wu, Y. Mallet, B. Walczak, W. Penninckx, D. L. Massart, S. Heuerding, and F. Erni. Comparison of regularized discriminant analysis, linear discriminant analysis and quadratic discriminant analysis, applied to NIR data. *Analytica Chimica Acta*, 329(3):257–265, 1996.

[44] Z. Xiaobo, Z. Jiewen, M. J. Povey, M. Holmes, and M. Hanpin. Variables selection methods in near-infrared spectroscopy, 2010.

[45] H. Zhao, K.-W. Huan, X.-G. Shi, F. Zhen, L.-Y. Liu, W. Liu, and C.-Y. Zhao. A Variable Selection Method of Near Infrared Spectroscopy Based on Automatic Weighting Variable Combination Population Analysis. *Chinese Journal of Analytical Chemistry*, 46(1):136–142, jan 2018.

[46] X. Zou, J. Zhao, and Y. Li. Selection of the efficient wavelength regions in FT-NIR spectroscopy for determination of SSC of 'Fuji' apple based on BiPLS and FiPLS models. *Vibrational Spectroscopy*, 44(2):220–227, jul 2007.