

Robust variable selection for model-based learning in presence of adulteration

Andrea Cappozzo ^{*} Francesca Greselin^{*}
 Thomas Brendan Murphy [†]

Abstract

The problem of identifying the most discriminating features when performing supervised learning has been extensively investigated. In particular, several methods for variable selection in model-based classification have been proposed. Surprisingly, the impact of outliers and wrongly labeled units on the determination of relevant predictors has received far less attention, with almost no dedicated methodologies available in the literature. In the present paper, we introduce two robust variable selection approaches: one that embeds a robust classifier within a greedy-forward selection procedure and the other based on the theory of maximum likelihood estimation and irrelevance. The former recasts the feature identification as a model selection problem, while the latter regards the relevant subset as a model parameter to be estimated. The benefits of the proposed methods, in contrast with non-robust solutions, are assessed via an experiment on synthetic data. An application to a high-dimensional classification problem of contaminated spectroscopic data concludes the paper.

1 Introduction

Nowadays, in many scientific domains such as chemometrics, computer vision, engineering and genetics among others, it is increasingly common to measure hundreds or thousands of variables on each sample. In principle, depending on the problem at hand, all the available features might be relevant and thus deemed to be included in a subsequent analysis. Most often, however, incorporating every piece of information at our disposal unnecessarily increases model complexity and, ultimately, it may undermine the entire output of a statistical procedure. Model-based methods are particularly sensitive to the well-known *curse of dimensionality* (Bellman, 1957), as such models are over-parametrized and suffer from identifiability problems in high dimensional spaces (Bouveyron

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

[†]School of Mathematics & Statistics and Insight Research Centre, University College Dublin, brendan.murphy@ucd.ie

and Brunet-Saumard, 2014; Bouveyron et al., 2019, Chapter 8). Therefore, in a discriminant analysis context, selecting the useful variables that better unveil the group structure is crucial to learn an efficient classifier. This has been known for a long time, as demonstrated by the specific literature reviews on the topic in the fields of machine learning (Blum and Langley, 1997; Yu and Liu, 2004; Liu and Motoda, 2007), data mining (Dash and Liu, 1997; Kohavi and John, 1997), bioinformatics (Saeys et al., 2007), genomic (Yu, 2008) and statistics (McLachlan, 1992; Guyon et al., 2007; Fop and Murphy, 2018). Nonetheless, the impact that outliers and wrongly labeled units cause on the efficient determination of discriminant variables has received far less attention. Indeed, the presence of attribute and class noise can heavily damage a classifier performance (Zhu and Wu, 2004), and most variable selection methods rely on the implicit assumption of dealing with an uncontaminated training set.

In order to overcome this limitation, the present paper proposes two approaches for robust variable selection in model-based classification: one that embeds a robust classifier, recently introduced in the literature, in a greedy-forward stepwise procedure for model selection (Section 4.1); and the other based on the theory of maximum likelihood and the notion of irrelevant variables within robust ML estimation of normal mixtures (Section 4.2). Both procedures rely on impartial trimming (Gordaliza, 1991): an appealing technique for robust parameter estimation in which no model assumption is a-priori required for the noise component. By leaving the anomalous units unmodeled, great flexibility is achieved and thus very heterogeneous contamination patterns can be effectively dealt with.

The remaining of the article is structured as follows. Section 2 formally characterizes the problem of variable selection in model-based discriminant analysis. In Section 3, the main features of the Robust Eigenvalue Decomposition Discriminant Analysis (REDDA) are reviewed. Two novel variable selection techniques resistant to outliers and label noise are introduced in Section 4: they are the main contributions of the present manuscript. Section 5 is devoted to the comparison of several feature selection procedures within two simulation studies in an artificially contaminated scenario. Section 6 presents a high-dimensional discrimination study where our proposals for robust variable selection are successfully applied to a chemometrics contest. Section 7 concludes the paper outlying some remarks and future research directions. Technical issues and computational details for the two novel methods are respectively deferred to A and B.

2 The problem of feature selection in discriminant analysis

The detection of p relevant features (out of the whole collection of $P \gg p$ available variables) on which to train the classifier is particularly desirable, as (McLachlan, 1992):

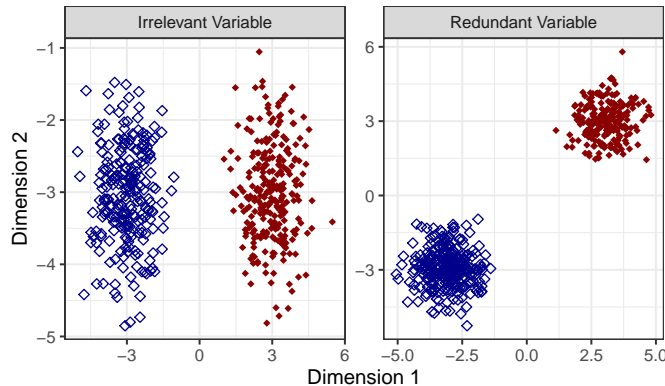


Figure 1: Examples of learning scenarios for which the second dimension is irrelevant (left panel) or redundant (right panel) in discriminating the two groups.

- it simplifies parameter estimation and interpretation;
- it avoids loss on predictive power due to the inclusion of irrelevant and redundant information;
- it leads to cost reduction on future data collection and processing.

Therefore, with the aim of choosing the best predictors, it is crucial to define the concept of “relevant variable”. The framework of model-based discriminant analysis allows to define “relevance” in terms of probabilistic dependence (or independence) with respect to the class membership (Ritter, 2014). The distribution of the *relevant variables*, i.e., features that bring significant information on class separation, directly depends on the class membership itself. In discriminating men and women of the same ethnicity for example, the height is naturally relevant. *Irrelevant or noisy variables*, on the contrary, do not contain any discriminating power, and hence their distribution is completely independent from the group structure. To continue with our previous example, hair and eye color do not convey any information on the gender of a person. 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. If the height of a person is known, little extra information is gained by finding out his/her head circumference for determining his/her gender. In Figure 1, the first dimension is a relevant variable for discriminating the two groups, while the second dimension is respectively irrelevant in the left panel and redundant in the right one.

Depending on how the variable selection process interacts with the model estimation, two general approaches for feature identification can be defined. Following the nomenclature introduced by John et al. (1994), *filter methods* are those in which the selection acts as a pre (or post) processing step, discarding variables whose distribution appears non-informative. Since the selection via

filter methods is performed separately from the model estimation, i.e., without reference to the class membership, such techniques may miss important grouping information; a standard example being Principal Component Analysis (Chang, 1983). For a state-of-the-art benchmark study on the comparison of filter methods for feature selection in high dimensional classification, the reader is referred to Bommert et al. (2020).

For the second class of methods the feature identification is “wrapped” around the classification procedure; hence they are denoted as *wrapper approaches*. Within this framework, variable selection and model estimation are simultaneously performed, aiming at identifying the predictors that better describe the underlying data partition. Focusing on the model-based methods for classification, Murphy et al. (2010) provide a wrapper approach for feature selection in semi-supervised discriminant analysis, recasting the feature identification as a model selection problem. The authors develop a greedy search and a head-long search algorithm for finding a local optimum in the model space, inspired by the seminal work on variable selection in model-based clustering of Dean et al. (2006), wherein for the first time the potential correlation between relevant and irrelevant variables is taken into account. Similarly, a general methodology for selecting predictors in model-based discriminant analysis is introduced in Maugis et al. (2011), where also theoretical results on model identifiability and consistency of the proposed criterion are validated. More recently, a regularization approach for feature selection in model-based clustering and classification is introduced in Celeux et al. (2019), where a lasso-like procedure is employed for overcoming the slowness yielded by stepwise algorithms when dealing with high-dimensional problems. The `SelvarMix` R package provides an efficient C++ implementation of the afore-mentioned procedure. Unfortunately, no one of the wrapper methods listed here provide protection against outliers and label noise: the presence of only few adulterated data points can severely undermine the variable selection results (see Section 5).

Lastly, methods that lie in between the two approaches have also been developed in the literature. Such hybrid methods usually involve feature selection based on some measure of separability between groups, like the one introduced by Indahl and Næs (2004), specifically tailored for spectroscopic data, and the one proposed by Andrews and McNicholas (2014). Further, a series of techniques based on metaheuristic strategies for variable selection in discriminant analysis can be found in Pacheco et al. (2006), while the method of Chiang and Pell (2004) relies on a stochastic search based on genetic algorithms. In general, even though being more complex and computationally intensive, wrapper approaches provide better classification results and more accurate representation of the data generating process (Kohavi and John, 1997). For this reason, the present manuscript will focus on wrapper approaches: the novel methods introduced in Section 4 fall within this category.

An important consideration to be made regards existing approaches that already provide robust selection of variables. In linear discriminant analysis (LDA), early-stage wrapper methods consider the employment of stepwise procedures in testing for no additional information, like the stepwise MANOVA de-

scribed in Section 12.3 of McLachlan (1992): these are usually based on the likelihood ratio test Wilks' Λ statistic. By respectively employing M-estimates and MCD-estimates to obtain a robust version of the Wilks' Λ statistics, Krusińska and Liebhart (1988) and Todorov (2007) develop LDA-based techniques for variable selection resistant to outliers. Nevertheless, to our best knowledge, wrapper methods that perform robust feature selection in a more general framework are still missing in the literature.

Prior to present our novel contributions for variable selection resistant to outliers and label noise, the Robust Eigenvalue Decomposition Discriminant Analysis (REDDA) model is briefly reviewed in the upcoming Section; for a thorough treatment the interested reader is referred to Cappozzo et al. (2020).

3 Robust model-based discriminant analysis

Model-based discriminant analysis (McLachlan, 1992; Fraley and Raftery, 2002) is a probabilistic framework for supervised classification, in which a classifier is built from a complete set of N learning observations (i.e., the training set):

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

where \mathbf{x}_n is 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, $g = 1, \dots, G$, with, $\sum_{g=1}^G l_{ng} = 1 \forall n = 1, \dots, N$. Alternatively, for sake of brevity, we will also employ the notation $l_n = g$ to denote the class of the n -th observation. We assume that the prior probability of group g is $\mathbb{P}(\mathbf{l} = g) = \tau_g$, with $\tau_g > 0$ and $\sum_{g=1}^G \tau_g = 1$. The g th class-conditional density is modeled with a P -dimensional Gaussian distribution with mean vector $\boldsymbol{\mu}_g \in \mathbb{R}^P$ and positive semi-definite covariance matrix $\boldsymbol{\Sigma}_g \in PD(P)$: $\mathbf{x}_n | \mathbf{l}_n = g \sim N_P(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$. Therefore, the joint density of $(\mathbf{x}_n, \mathbf{l}_n)$ is given by:

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

where $\phi(\cdot; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ denotes the multivariate normal density and $\boldsymbol{\theta}$ is the collection of parameters to be estimated, $\boldsymbol{\theta} = \{\tau_1, \dots, \tau_G, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_G\}$. Eigenvalue Decomposition Discriminant Analysis (EDDA) is a family of classifiers developed from the probabilistic structure in (2), wherein different assumptions about the covariance matrices are considered. Particularly, EDDA is based on the following eigenvalue decomposition (Banfield and Raftery, 1993; Celeux and Govaert, 1995):

$$\boldsymbol{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g' \quad (3)$$

where \mathbf{D}_g is an orthogonal matrix of eigenvectors, \mathbf{A}_g is a diagonal matrix such that $|\mathbf{A}_g| = 1$ and $\lambda_g = |\boldsymbol{\Sigma}_g|^{1/p}$. These elements correspond respectively to

the orientation, shape and volume (alternatively called scale) of the Gaussian components. Allowing each parameter in (3) to be equal or different across groups, Bensmail and Celeux (1996) defined a family of 14 patterned models. Capozzo et al. (2020) introduced a robust modification to EDDA, hereafter denoted REDDA, in which parameter estimates are protected against label noise and outliers by means of a *trimmed mixture log-likelihood* (Neykov et al., 2007):

$$\ell_{trim}(\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{X}, \mathbf{l}) = \sum_{n=1}^N \zeta(\mathbf{x}_n) \sum_{g=1}^G l_{ng} \log(\tau_g \phi(\mathbf{x}_n; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)) \quad (4)$$

where $\zeta(\cdot)$ is a 0-1 trimming indicator function, that expresses whether observation \mathbf{x}_n is trimmed off or not. A fixed fraction γ of observations is unassigned by setting $\sum_{n=1}^N \zeta(\mathbf{x}_n) = \lceil N(1-\gamma) \rceil$. The *labelled trimming level* γ accounts for possible adulteration, namely outliers and label noise, in the training set. Maximization of (4) is carried out via a generalization of the FastMCD algorithm by Rousseeuw and Driessen (1999), adapted to deal with parsimonious structures in the covariance matrices. Particularly, in this context the *Concentration step* (C-step) is enforced by temporarily discarding $\lfloor N\gamma \rfloor$ units with lowest value of:

$$\phi(\mathbf{x}_n; \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g) \quad g = 1, \dots, G. \quad (5)$$

For these observations, $\zeta(\mathbf{x}_n) = 0$ in (4) as they will not be accounted for in the next estimation step: the algorithm stops once the less plausible $\lfloor N\gamma \rfloor$ discarded units, out of the N units in the learning set, are confirmed to be the same on two consecutive iterations. Notice that the mixing proportions $\hat{\tau}_g$ do not appear in (5): the estimated group-conditional densities act as discriminative tools for trimming, so that $\lfloor N\gamma \rfloor$ overall samples are removed at each step. At the end of the procedure, a value of $\zeta(\mathbf{x}_n) = 0$ corresponds to identify \mathbf{x}_n as an unreliable unit. The REDDA classifier can then be employed for assigning an unlabeled sample \mathbf{y}_m , $m = 1, \dots, M$ (i.e., the test set) to the class g whose associated posterior probability

$$\hat{z}_{mg} = \frac{\hat{\tau}_g \phi(\mathbf{y}_m; \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g)}{\sum_{j=1}^G \hat{\tau}_j \phi(\mathbf{y}_m; \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)}. \quad (6)$$

is highest, by means of the usual maximum a posteriori (MAP) rule. In addition, also the trimmed units can be a-posteriori assigned to the component g displaying the highest value of $\hat{\tau}_g \phi(\mathbf{x}_n; \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g)$, to recover a reasonable label for observations that previously got an adulterated one.

4 Robust Variable Selection in model-based classification

In the present Section we introduce two novel wrapper approaches for robust variable selection in high-dimensional model-based classification.

In Section 4.1, the REDDA method is embedded in a greedy-forward procedure for model selection. A robust classification rule is constructed in a step-wise manner, by considering the inclusion of extra variables and also the removal of existing variables to/from the model, conditioning on their discriminating power. Particularly, the selection procedure is based on a robust information criterion, that accounts for the possible presence of outliers and label noise in the dataset.

In Section 4.2, the theory of maximum likelihood estimation and the notion of irrelevant variables for normal mixtures is employed for defining a ML subset selector, along the lines of the procedure introduced in section 5.3.3 of Ritter (2014) for the unsupervised framework. The identification of the relevant subset is regarded as a parameter to be estimated via ML: an algorithmic procedure is derived for maximizing the objective function. The Section concludes with a comparison, highlighting strengths and weaknesses of the two proposals.

4.1 The robust stepwise greedy-forward approach via TBIC

The present procedure searches for the set of relevant variables in a greedy-stepwise manner. That is, we start from the empty set and we sequentially add relevant variables until no more discriminating features are available. More specifically, following the notation introduced in Section 3, in each step of the algorithm we partition the learning observations \mathbf{x}_n , $n = 1, \dots, N$, 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.

In order to decide whether to include the proposed variable x_n^p , we compare 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}):

$$p(\mathbf{x}_n | \mathbf{l}_n) = p(\mathbf{x}_n^c, x_n^p, \mathbf{x}_n^o | \mathbf{l}_n) = p(\mathbf{x}_n^c | \mathbf{l}_n) p(x_n^p | \mathbf{x}_n^r \subseteq \mathbf{x}_n^c) p(\mathbf{x}_n^o | x_n^p, \mathbf{x}_n^c)$$

where \mathbf{x}_n^r denotes a subset of the currently included variables \mathbf{x}_n^c . The grouping model specifies that x_n^p provides extra grouping information beyond that provided by \mathbf{x}_n^c ; whereas the No Grouping model specifies that x_n^p is conditionally independent of the group membership given \mathbf{x}_n^r . The reason for considering \mathbf{x}_n^r in the conditional distribution being that x_n^p might be related to only a subset of the grouping variables \mathbf{x}_n^c (Maugis et al., 2009a,b, 2011). The differences between the two models are graphically illustrated in Figure 2. The model

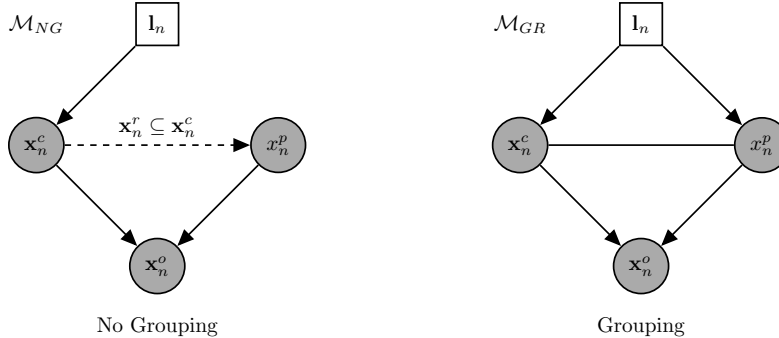


Figure 2: Graphical Representation of the Grouping and the No Grouping models

structure of $p(\mathbf{x}_n^o | x_n^p, \mathbf{x}_n^c)$ is assumed to be the same for both grouping and no grouping specification, and we let $p(\mathbf{x}_n^c, x_n^p | \mathbf{l}_n)$ and $p(\mathbf{x}_n^c | \mathbf{l}_n)$ be a normal density with parsimonious covariance structure, according to the model assumptions introduced in the previous Section. Additionally, we assume $p(x_n^p | \mathbf{x}_n^c \subseteq \mathbf{x}_n^c)$ to be a normal linear regression model, as a result from conditional multivariate normal means. The selection of which model to prefer is carried out employing a robust approximation to the Bayes Factor. More specifically, the Bayes Factor (Kass and Raftery, 1995) is equal to the ratio between the integrated likelihood of the two competing models:

$$\mathcal{B}_{GR,NG} = \frac{p(\mathbf{x}_n | \mathcal{M}_{GR})}{p(\mathbf{x}_n | \mathcal{M}_{NG})} = \frac{\int p(\mathbf{x}_n | \boldsymbol{\theta}_{GR}, \mathcal{M}_{GR}) p(\boldsymbol{\theta}_{GR} | \mathcal{M}_{GR}) d\boldsymbol{\theta}_{GR}}{\int p(\mathbf{x}_n | \boldsymbol{\theta}_{NG}, \mathcal{M}_{NG}) p(\boldsymbol{\theta}_{NG} | \mathcal{M}_{NG}) d\boldsymbol{\theta}_{NG}} \quad (7)$$

where $\boldsymbol{\theta}_{GR}$ and $\boldsymbol{\theta}_{NG}$ denote the set of parameters for the Grouping (GR) and the No Grouping (NG) model, respectively. When no prior preference for one of the two models is considered, (7) is equal to the posterior odds in favour of \mathcal{M}_{GR} . The Bayes Factor can therefore be used for assessing to which extent the data supports the GR structure compared to the NG formulation. Along the lines of Raftery and Dean (2006), the Bayesian Information Criterion

$$BIC = 2 \times \log \text{maximized likelihood} - v \log N$$

is used as an approximation for the integrated likelihood, where v is a penalty term (number of parameters in the model) and N is the sample size (Schwarz, 1978). Thus, twice the logarithm of $\mathcal{B}_{GR,NG}$ can be approximated with

$$2 \log (\mathcal{B}_{GR,NG}) \approx BIC(GR) - BIC(NG) \quad (8)$$

and a variable x_n^p with a positive difference in $BIC(GR) - BIC(NG)$ is a candidate for being added to the model. For avoiding the detrimental effect that class and attribute noise might produce in the variable selection procedure, the

Trimmed BIC (TBIC), firstly introduced in Neykov et al. (2007), is employed as a robust proxy for the quantities in (8). Let us define:

$$TBIC(GR) = 2 \underbrace{\sum_{n=1}^N \zeta(\mathbf{x}_n^c, x_n^p) \sum_{g=1}^G l_{ng} \log \left(\hat{\tau}_g^{cp} \phi(\mathbf{x}_n^c, x_n^p; \hat{\boldsymbol{\mu}}_g^{cp}, \hat{\boldsymbol{\Sigma}}_g^{cp}) \right)}_{2 \times \text{trimmed log maximized likelihood of } p(\mathbf{x}_n^c, x_n^p, \mathbf{I}_n)} - v^{cp} \log(N^*) \quad (9)$$

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

$$+ 2 \underbrace{\sum_{n=1}^N \iota(\mathbf{x}_n^c, x_n^p) \log \left[\phi \left(x_n^p; \hat{\alpha} + \hat{\boldsymbol{\beta}}' \mathbf{x}_n^r, \hat{\sigma}^2 \right) \right]}_{2 \times \text{trimmed log maximized likelihood of } p(x_n^p | \mathbf{x}_n^r \subseteq \mathbf{x}_n^c)} - v^p \log(N^*). \quad (10)$$

The penalty terms v^{cp} and v^c indicate the number of parameters for a REDDA model respectively estimated on the set of variables \mathbf{x}_n^c, x_n^p and \mathbf{x}_n^c ; while v^p accounts for the number of parameters in the linear regression of x_n^p on \mathbf{x}_n^r . The 0-1 indicator functions $\zeta(\cdot)$ and $\iota(\cdot)$ identify the subset of observations that have null weight in the trimmed likelihood under the grouping and no grouping models, with $N^* = \sum_{n=1}^N \zeta(\mathbf{x}_n) = \sum_{n=1}^N \iota(\mathbf{x}_n)$.

In detail, the parameters $\{\tau_g^{cp}, \boldsymbol{\mu}_g^{cp}, \boldsymbol{\Sigma}_g^{cp}\}, g = 1, \dots, G$ of the grouping model are estimated through a standard REDDA fitted on the variables \mathbf{x}_n^c, x_n^p , in which the C-step is enforced discarding the $\lfloor N\gamma \rfloor$ samples with lowest value of

$$D_{Grouping}(\mathbf{x}_n^c, x_n^p; \hat{\boldsymbol{\theta}}_{GR}) = \sum_{g=1}^G l_{ng} \log \left[\phi \left(\mathbf{x}_n^c, x_n^p; \hat{\boldsymbol{\mu}}_g^{cp}, \hat{\boldsymbol{\Sigma}}_g^{cp} \right) \right] \quad n = 1, \dots, N \quad (11)$$

likewise for the general case in (5). For the no grouping model, REDDA needs to be fitted only on the set of currently included variables \mathbf{x}_n^c , coupled with the linear regression of x_n^p on \mathbf{x}_n^r . For this case, the discriminating function reads:

$$D_{No\ Grouping}(\mathbf{x}_n^c, x_n^p; \hat{\boldsymbol{\theta}}_{NG}) = \sum_{g=1}^G l_{ng} \log \left[\phi \left(\mathbf{x}_n^c; \hat{\boldsymbol{\mu}}_g^c, \hat{\boldsymbol{\Sigma}}_g^c \right) \right] + \log \left[\phi \left(x_n^p; \hat{\alpha} + \hat{\boldsymbol{\beta}}' \mathbf{x}_n^r, \hat{\sigma}^2 \right) \right] \quad (12)$$

for $n = 1, \dots, N$. That is, at each iteration of the procedure that leads to the final robust estimates, we discard $\lfloor N\gamma \rfloor$ samples with the lowest contribution to the conditional likelihood under the no grouping model. Once the C-step is enforced, the set of parameters $\{\alpha, \boldsymbol{\beta}, \sigma^2\}$ for the regression part is robustly estimated via ML on the untrimmed observations, in which a stepwise method

is employed for automatically choosing the subset of regressors \mathbf{x}_n^r . Further details concerning the implementation are included in A.

After each addition stage, we make use of the same procedure described above to check whether an already chosen variable in \mathbf{x}_n^c should be removed: in this case x_n^p takes the role of the variable to be dropped, and a negative difference in terms of TBIC implies the exclusion of x_n^p to the set of currently included variables. The procedure iterates between variable addition and removal stage until two consecutive steps have been rejected, then it stops. Notice that, whenever $\gamma = 0$, BIC and TBIC coincide and the entire approach reduces to the methodology described in Maugis et al. (2011).

A last worthy note regards the theoretical justification for the employment of TBIC as an approximation of the integrated likelihood. The rationale arises from the spurious outliers model, firstly introduced in Gallegos and Ritter (2005), as the probabilistic specification for the contaminated sub-sample. Let q_n denote an indicator of genuine observations, such that $q_n = 1$ when $\{(\mathbf{x}_n, \mathbf{l}_n)\}$ is a “regular” unit and $q_n = 0$ whenever $\{(\mathbf{x}_n, \mathbf{l}_n)\}$ presents some sort of contamination/adulteration. Notice that the complete observation $\{(\mathbf{x}_n, \mathbf{l}_n)\}$ might be regarded as an outlier whenever either the associated label and/or some of its predictors present unusual values. In such a way, we account for both attribute and class noise. The data generating distribution for a specific observation $\{(\mathbf{x}_n, \mathbf{l}_n)\}$ is then assumed to be as follows:

$$p(\mathbf{x}_n, \mathbf{l}_n | q_n; \boldsymbol{\theta}) = p(\mathbf{x}_n, \mathbf{l}_n; \boldsymbol{\theta})^{q_n} w(\mathbf{x}_n, \mathbf{l}_n; \boldsymbol{\psi}_n)^{(1-q_n)} \quad (13)$$

where $p(\mathbf{x}_n, \mathbf{l}_n; \boldsymbol{\theta})$ denotes the probability distribution for the regular bulk of the data, in our context being alternatively the Grouping or the No Grouping model; and $w(\mathbf{x}_n, \mathbf{l}_n; \boldsymbol{\psi}_n)$ is an almost arbitrary, subject specific probability density function, parametrized by $\boldsymbol{\psi}_n \in \boldsymbol{\Psi}_n$. For an independent sample of N observations, the likelihood for the model in (13) is therefore given by:

$$\prod_{n=1}^N p(\mathbf{x}_n, \mathbf{l}_n; \boldsymbol{\theta})^{q_n} \prod_{n=1}^N w(\mathbf{x}_n, \mathbf{l}_n; \boldsymbol{\psi}_n)^{(1-q_n)} \quad (14)$$

where a fixed $\gamma\%$ of contamination is assumed such that $N^* = \sum_{n=1}^N q_n = \lceil N(1 - \gamma) \rceil$.

Let $\mathcal{N} = \{N_1, N_0\}$ be a partition of N into regular and non-regular observations, indexed by q_n being either 1 or 0 for $n = 1, \dots, N$, with $|N_1| = \lceil N(1 - \gamma) \rceil$ and $|N_0| = \lfloor N\gamma \rfloor$, respectively. Further, denote with $\mathcal{D}(N)$ the set of all partitions of such type, with $|\mathcal{D}(N)| = \binom{N}{\lceil N(1 - \gamma) \rceil}$. The non-regular contribution of the contaminated observations can be avoided in maximizing (14) with respect to $\boldsymbol{\theta}$ when the $w(\cdot; \boldsymbol{\psi}_n)$ s satisfy

$$\arg \max_{\mathcal{N} \in \mathcal{D}(N)} \max_{\boldsymbol{\theta}} \prod_{n=1}^N p(\mathbf{x}_n, \mathbf{l}_n; \boldsymbol{\theta})^{q_n} \subseteq \arg \max_{\mathcal{N} \in \mathcal{D}(N)} \max_{\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_N} \prod_{n=1}^N w(\mathbf{x}_n, \mathbf{l}_n; \boldsymbol{\psi}_n)^{(1-q_n)}. \quad (15)$$

The condition in (15) means that the configuration that maximizes the first term in (14) automatically maximizes the second one (Gallegos and Ritter, 2005). More specifically, the partitions assigning $\lceil N(1 - \gamma) \rceil$ regular units that maximize the likelihood of the genuine observations are contained in the set of partitions assigning $\lfloor N\gamma \rfloor$ non regular units that maximize the likelihood corresponding to the noise. Condition (15) holds under general and non-restrictive assumptions on the non regular units, particularly, $w(\cdot; \boldsymbol{\psi}_n)$ can easily accommodate observations that can be merely regarded as outliers (Gallegos and Ritter, 2005; García-Escudero et al., 2008). The contaminated observations are therefore no more considered in the estimation process, and the model log-likelihood simplifies to:

$$\sum_{n=1}^N q_n \log p(\mathbf{x}_n, \mathbf{l}_n; \boldsymbol{\theta}) \quad (16)$$

to be maximized with respect to the set of parameters $\boldsymbol{\theta}$; details are reported in A. Finally, the integrated log-likelihood for (16) can be approximated via the Bayesian Information Criterion:

$$2 \sum_{n=1}^N q_n \log p(\mathbf{x}_n, \mathbf{l}_n; \hat{\boldsymbol{\theta}}) - v \log N^* \quad (17)$$

where $\hat{\boldsymbol{\theta}}$ denotes MLE for the simplified log-likelihood, v is the number of parameters and N^* is the number of data values that contribute to the summation in (16) (Kass, 1993). Depending which scenario is considered, (17) defines (9) or (10) under the Grouping and the No Grouping model, respectively.

4.2 The ML subset selector approach

The second approach we consider for robust variable selection in model-based classification stems from the maximum likelihood subset selector theory developed for clustering, where the main reference is Section 5.3.3 of Ritter (2014). Particularly, being classification a generally simpler problem than unsupervised learning, the ML subset selection ideas are naturally adapted to a robust supervised context with variable selection. Here we build a model for the entire P -dimensional space in which the observations lie, exploiting theoretical results for the conditional distribution of the multivariate Gaussian under irrelevance. Let us introduce the following notation: for $\boldsymbol{\Sigma} \in PD(P)$, denote its restriction to the variables in $F \subseteq 1, \dots, P$ by $\boldsymbol{\Sigma}_F$, with size $|F| = p$. The block-wise representation of $\boldsymbol{\Sigma}$, via the natural order of F , is therefore:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_F & \boldsymbol{\Sigma}_{F,E} \\ \boldsymbol{\Sigma}_{E,F} & \boldsymbol{\Sigma}_E \end{pmatrix}$$

with $E = \bar{F}$ and $|E| = P - p$. Analogously, the vector $\boldsymbol{\mu}_F$ is the projection of $\boldsymbol{\mu} \in \mathbb{R}^P$ onto the variables in F , following the natural order of F . For a generic observation $\mathbf{x}_n \in \mathbb{R}^P$, the canonical projection of a normal distribution

to a subset F of variables is described by the restrictions $\boldsymbol{\mu}_F$ and $\boldsymbol{\Sigma}_F$ of its parameters, with the equality $N_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}(\mathbf{x}_{n,F}) = N_{\boldsymbol{\mu}_F, \boldsymbol{\Sigma}_F}(\mathbf{x}_{n,F})$ such that $\mathbf{x}_{n,F} \sim N(\boldsymbol{\mu}_F, \boldsymbol{\Sigma}_F)$. Considering the notation introduced in Section 3 and applying standard results for multivariate normal theory, (see, for example, Theorem 3.2.4 in Mardia et al. (1979)), the conditional distribution of $\mathbf{x}_{n,E}$ given $\mathbf{x}_{n,F}, \mathbf{l}_n$ reads:

$$\mathbf{x}_{n,E} | \mathbf{x}_{n,F}, \mathbf{l}_n = g \sim \mathcal{N}_{P-p}(\boldsymbol{\mu}_{g,E|F} + \mathbf{G}_{g,E|F} \mathbf{x}_{n,F}; \boldsymbol{\Sigma}_{g,E|F}) \quad (18)$$

where $\boldsymbol{\mu}_{g,E|F} = \boldsymbol{\mu}_{g,E} - \mathbf{G}_{g,E|F} \boldsymbol{\mu}_{g,F}$, $\boldsymbol{\Sigma}_{g,E|F} = \boldsymbol{\Sigma}_{g,E} - \mathbf{G}_{g,E|F} \boldsymbol{\Sigma}_{g,F} \mathbf{G}_{g,E|F}$ and $\mathbf{G}_{g,E|F} = \boldsymbol{\Sigma}_{g,E,F} \boldsymbol{\Sigma}_{g,F}^{-1}$, $g = 1, \dots, G$. Now assume that E is an irrelevant subset with respect to F , that is, the class membership \mathbf{l}_n is conditionally independent of $\mathbf{x}_{n,E}$ given $\mathbf{x}_{n,F}$. By Lemma 5.2 and Theorem 5.7 of Ritter (2014), the parameters $\mathbf{G}_{g,E|F}$, $\boldsymbol{\mu}_{g,E|F}$ and $\boldsymbol{\Sigma}_{g,E|F}$ do not depend on class g ; applying the product formula we thus obtain the following specification for the joint density of $(\mathbf{x}_{n,F}, \mathbf{x}_{n,E}, \mathbf{l}_n)$:

$$\begin{aligned} p(\mathbf{x}_{n,F}, \mathbf{x}_{n,E}, \mathbf{l}_n) &= p(\mathbf{x}_{n,F}, \mathbf{x}_{n,E} | \mathbf{l}_n) p(\mathbf{l}_n) = \\ &= p(\mathbf{x}_{n,E} | \mathbf{x}_{n,F}, \mathbf{l}_n) p(\mathbf{x}_{n,F} | \mathbf{l}_n) p(\mathbf{l}_n) = \\ &= p(\mathbf{x}_{n,E} | \mathbf{x}_{n,F}) p(\mathbf{x}_{n,F} | \mathbf{l}_n) p(\mathbf{l}_n). \end{aligned} \quad (19)$$

Therefore, for a sample of N observations, drawn from the random variable \mathbf{X} , the associated trimmed log-likelihood for the probability density in (19) is:

$$\begin{aligned} \ell_{trim}(\boldsymbol{\tau}, \boldsymbol{\mu}_F, \boldsymbol{\Sigma}_F, \mathbf{G}_{E|F}, \boldsymbol{\mu}_{E|F}, \boldsymbol{\Sigma}_{E|F} | \mathbf{X}, \mathbf{l}) &= \\ &= \sum_{n=1}^N \zeta(\mathbf{x}_n) \left(\sum_{g=1}^G l_{ng} \log [\tau_g \phi(\mathbf{x}_{n,F}; \boldsymbol{\mu}_{g,F}, \boldsymbol{\Sigma}_{g,F})] + \right. \\ &\quad \left. + \log [\phi(\mathbf{x}_{n,E} - \mathbf{G}_{E|F} \mathbf{x}_{n,F}; \boldsymbol{\mu}_{E|F}, \boldsymbol{\Sigma}_{E|F})] \right) \end{aligned} \quad (20)$$

where the identification of the relevant variables belonging to the subset F is regarded as a model parameter. Maximization of (20) is carried out via a modification of the EMST algorithm introduced in Ritter (2014), adapted to the classification framework and extended to flexibly account for the entire family of patterned models of Bensmail and Celeux (1996). The main steps involving the estimation procedure are given below, further details concerning the implementation can be found in B.

1. Robust Initialization:

- If N is sufficiently large compared to P and G , draw a random $(P + 1)$ -subset for each class g , $g = 1, \dots, G$. The first M-step will be computed only on such units: this is achieved by setting $\zeta(\mathbf{x}_n) = 1$ if \mathbf{x}_n belongs to the drawn subset, otherwise $\zeta(\mathbf{x}_n) = 0$. Go to step 2 of the algorithm.

- If N is small compared to P and G , draw a random $(p + 1)$ -subset for each class g , $g = 1, \dots, G$ and set $\zeta(\mathbf{x}_n) = 1$ if \mathbf{x}_n belongs to any of such G subsets, otherwise set $\zeta(\mathbf{x}_n) = 0$.

Draw a random subset $\hat{F}^{(0)}$ of dimension p from $1, \dots, P$ and compute:

$$\hat{\boldsymbol{\mu}}_{g, \hat{F}^{(0)}} = \frac{\sum_{n=1}^N \zeta(\mathbf{x}_n) l_{ng} \mathbf{x}_n}{\sum_{n=1}^N \zeta(\mathbf{x}_n) l_{ng}} \quad g = 1, \dots, G,$$

and $\hat{\boldsymbol{\Sigma}}_{g, \hat{F}^{(0)}}^{(0)}$, $g = 1, \dots, G$, depending on the considered patterned model, refer to Bensmail and Celeux (1996) for the details. Lastly, update the trimming function $\zeta(\mathbf{x}_n)$, $n = 1, \dots, N$, setting $\zeta(\mathbf{x}_n) = 0$ for the $\lfloor N\gamma \rfloor$ samples with lowest value of

$$l_{ng} \log \left[\phi(\mathbf{x}_n, F^{(0)}; \hat{\boldsymbol{\mu}}_{g, F^{(0)}}^{(0)}, \hat{\boldsymbol{\Sigma}}_{g, F^{(0)}}^{(0)}) \right]$$

and $\zeta(\mathbf{x}_n) = 1$ otherwise.

2. (*M-step*)

Compute:

$$\hat{\tau}_g = \frac{\sum_{n=1}^N \zeta(\mathbf{x}_n) l_{ng}}{\lfloor N(1 - \gamma) \rfloor} \quad g = 1, \dots, G$$

$$\hat{\boldsymbol{\mu}}_g = \frac{\sum_{n=1}^N \zeta(\mathbf{x}_n) l_{ng} \mathbf{x}_n}{\sum_{n=1}^N \zeta(\mathbf{x}_n) l_{ng}} \quad g = 1, \dots, G.$$

Estimation of $\boldsymbol{\Sigma}_g$ depends on the considered patterned model, details are given in Bensmail and Celeux (1996).

Notice that the estimates are computed for the full dimension P , that is $\hat{\boldsymbol{\mu}}_g \in \mathbb{R}^P$ and $\hat{\boldsymbol{\Sigma}}_g \in PD(P)$, respectively. In addition, robustly compute also the pooled mean:

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{n=1}^N \zeta(\mathbf{x}_n) \mathbf{x}_n}{\lfloor N(1 - \gamma) \rfloor}.$$

Depending on the considered patterned model, formulae for the associated pooled estimate $\hat{\boldsymbol{\Sigma}}$ are detailed in B.

3. (*S-step*)

Minimize the difference:

$$h(F) = \sum_{g=1}^G \hat{\tau}_g \log \det \hat{\boldsymbol{\Sigma}}_{g, F} - \log \det \hat{\boldsymbol{\Sigma}}_F \quad (21)$$

w.r.t. the subset $\hat{F} \subseteq 1, \dots, P$, with $|\hat{F}| = p$, where $\hat{\boldsymbol{\Sigma}}_{g, \hat{F}}$ is the restriction of $\hat{\boldsymbol{\Sigma}}_g$ to \hat{F} . The minimization of (21) involves a discrete structure optimization, that becomes quickly unfeasible as $\binom{P}{p}$ grows: a genetic algorithm is proposed for solving it (more details in B).

4. (*T-step*)

Compute the MLE's for the regression parameters

$$\begin{aligned}\hat{G}_{\hat{E}|\hat{F}} &= \hat{\Sigma}_{\hat{E},\hat{F}}\hat{\Sigma}_{\hat{F}}^{-1} \\ \hat{\mu}_{\hat{E}|\hat{F}} &= \hat{\mu}_{\hat{E}} - \hat{G}_{\hat{E}|\hat{F}}\hat{\mu}_{\hat{F}} \\ \hat{\Sigma}_{\hat{E}|\hat{F}} &= \hat{\Sigma}_{\hat{E}} - \hat{\Sigma}_{\hat{E},\hat{F}}\hat{\Sigma}_{\hat{F}}^{-1}\hat{\Sigma}_{\hat{F},\hat{E}}\end{aligned}$$

and update the value of the trimming function $\zeta(\cdot)$, setting $\zeta(\mathbf{x}_n) = 0$ for the $\lfloor N\gamma \rfloor$ samples with lowest value of

$$\sum_{g=1}^G l_{ng} \log \left[\hat{\tau}_g \phi(\mathbf{x}_{n,\hat{F}}; \hat{\mu}_{g,\hat{F}}, \hat{\Sigma}_{g,\hat{F}}) \right] + \log \left[\phi \left(\mathbf{x}_{n,\hat{E}} - \hat{G}_{\hat{E}|\hat{F}} \mathbf{x}_{n,\hat{F}}; \hat{\mu}_{\hat{E}|\hat{F}}, \hat{\Sigma}_{\hat{E}|\hat{F}} \right) \right].$$

5. Iterate 2 – 4 until the $\lfloor N\gamma \rfloor$ discarded observations are exactly the same on two consecutive iterations, then stop.

The procedure described in steps 1-5 shall be performed `n_init` times: the parameter estimates that lead to the highest value of the objective function (20), out of `n_init` repetitions, provide the final estimated quantities. As a last worthy comment, notice that the specification of the cardinality of F , i.e., the number p of relevant variables that are sought by the algorithm, is a-priori required as a model hyper-parameter.

4.3 Methods comparison

In the previous subsections two novel methods for robust variable selection in model-based classification have been introduced. As already anticipated, the main operational difference between the two relies on the fact that the ML subset selector requires the a-priori specification of the subset-size p , whereas the greedy-forward approach via TBIC automatically infers the number of relevant variables by means of a stopping criterion in the stepwise search. This could come both as an advantage and as a disadvantage: one may desire to specifically retain the p most relevant variables (i.e., $p = 2$ for visualization purposes). In this case, the ML subset selector approach shall be preferred, as the entire feature space P is accounted for in the likelihood specification in (20), contrarily to the greedy approach employed in Section 4.1. If this is not the case, run the algorithm for a reasonable range of values p and select the favourite solution, consensus methods like the one in Strehl and Ghosh (2002) for clustering can be adapted to the classification framework. In addition, if computational burden is not an issue, the greedy-forward approach via TBIC can be firstly employed for assessing the order of magnitude of the subset size, and afterwards the ML subset selector can be run varying p in the proximity of the number of relevant variables found by the former method, qualitatively assessing the difference.

Clearly, the suggestions above are mostly heuristic, a more formal treatment on how to compare and validate results from both procedures is still missing:

this however goes beyond the scope of the present manuscript and it will be the object of future research.

4.4 On the choice of the impartial trimming level γ

Both methodologies introduced in the previous sections require the a-priori specification of the impartial trimming level γ . Sensibly setting such a hyperparameter is not an easy task, as it implicitly requires the user to estimate the degree of contamination he/she is expected to find in a dataset. How to automatically infer, in a data-driven fashion, the true noise percentage present in the data is still an open issue in the robust clustering literature, even though several procedures have been recently proposed to mitigate and/or partially solve the problem: see for example García-Escudero et al. (2011); Dotto et al. (2018); Cerioli et al. (2018, 2019); Riani et al. (2019) and references therein. Given that, providing an exhaustive solution to this age-old problem goes beyond the scope of the present manuscript; nevertheless, some remarks and suggestions must be made on this regard.

Start by noticing that robust variable selection adds a layer of difficulty to the learning framework. More precisely, in this context we would firstly like to prevent adulterated units from jeopardizing the identification of relevant features, and, subsequently, to perform robust estimation and outlier detection on the subset of retained variables. While the two stages are clearly interconnected, we have observed that the performance of the former is generally less dependent by the choice of γ than the latter. Particularly, numerical experiments have highlighted that underestimating the true contamination level tends to favor the inclusion of several redundant and/or irrelevant variables among the relevant ones (see Section 5). On the other hand, the sensitivity study reported in Section 5.3 reveals that setting a higher than necessary trimming level has much less negative impact in the important features identification. Motivated by these arguments we suggest to, at least initially, set a precautionary high value of γ when it comes to variable selection, and, once the relevant subset has been identified, to tune it on the so-obtained lower dimensional feature space by exploiting the techniques already available in the literature. Operationally on the other hand, an adaptive procedure can be built by starting the search from a high level of trimming γ_{MAX} , and subsequently monitoring the changes in the retained variables subset for decreasing values of γ . This solution stems from the ideas discussed in Riani et al. (2019), where the partition stability of a robust clustering procedure is monitored by computing the Adjusted Rand Index (ARI; Rand, 1971) between consecutive allocations. In our context, whenever the magnitude and/or the configuration of two consecutive subsets of retained variables change, it can be interpreted as a sign that some contaminated units could have spoiled the procedure. Notice that standard information theoretic concepts, like the Hamming distance (Hamming, 1950), can be directly used to compare solutions obtained via the ML subset selector method: the relevant subset size p is a-priori set and assumed to be kept fixed during the search. On the other hand, more involved metrics, like, for example, Levenshtein dis-

tances (Dan, 1997), are needed to compare outputs from the stepwise approach via TBIC, as their magnitude may vary when different trimming values are considered.

As a last worthy note, we highlight the fact that our procedures are specifically designed to prevent attribute and class noise to deteriorate the variable selection process. Once this has been performed, extra effort shall be invested in effectively unraveling how many anomalous observations are contaminated within the set of retained features. That is, we are not directly interested in identifying outliers that can certainly arise in those dimensions that are not important for the classification task: a relevant example on this matter is showcased by sample 43 on the starches discrimination study in Section 6. To this extent, coupling cellwise outlier detection (Farcomeni, 2014; Rousseeuw and Bossche, 2018) and variable selection could be a promising attempt to cast light on this problem: it will be the object of future research.

5 Simulation studies

The aim of this Section is to numerically assess the effectiveness of the methodologies introduced in Section 4, whilst investigating the effect that contamination has on standard variable selection procedures. In doing so, we decided to rely on the same data generating process (DGP) considered in Maugis et al. (2011) and Celeux et al. (2019), described in Section 5.1, including in addition some attribute and class noise to the original experiment. Firstly, a simulated example with a fixed level of contamination is presented in Section 5.2, in which accuracy metrics are computed for both robust and non-robust variable selection methods. Secondly, a sensitivity study is reported in Section 5.3, displaying how different trimming and contamination levels affect the novel procedures.

5.1 Experimental Setup

The synthetic dataset considers $G = 4$ classes for a total of $P = 16$ features: the first three are relevant for the classification, the subsequent four are redundant given the first ones, while the last nine are independent from both the group variable and the previous predictors. The prior probabilities of the four classes are equal to $\boldsymbol{\tau} = (0.15, 0.3, 0.2, 0.35)$. On the three discriminant variables, data are generated from multivariate normal densities

$$\mathbf{x}_n^{[1-3]} | \mathbf{1}_n = g \sim \phi(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \quad g = 1, \dots, 4$$

with mean vectors

$$\begin{aligned} \boldsymbol{\mu}_1 &= (1.5, -1.5, 1.5)', & \boldsymbol{\mu}_2 &= (-1.5, 1.5, 1.5)' \\ \boldsymbol{\mu}_3 &= (1.5, -1.5, -1.5)', & \boldsymbol{\mu}_4 &= (-1.5, 1.5, -1.5)' \end{aligned}$$

and covariance matrices Σ_g with elements $\rho_g^{|i-j|}$, $1 \leq i, j \leq 3$, and $\rho_1 = 0.85$, $\rho_2 = 0.1$, $\rho_3 = 0.65$, $\rho_4 = 0.5$. The four redundant variables are sampled from

$$\mathbf{x}_n^{[4-7]} \sim N\left(\mathbf{x}_n^{[1,3]} \mathbf{B}; \mathbf{I}_4\right), \quad \mathbf{B} = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & -2 & 2 & 1 \end{pmatrix},$$

while the 9 independent ones are simulated from $\mathbf{x}_n^{[8-16]} \sim N(\boldsymbol{\eta}, \boldsymbol{\delta})$ with

$$\boldsymbol{\eta} = (-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2)$$

and

$$\boldsymbol{\delta} = \text{diag}(0.5, 0.75, 1, 1.25, 1.5, 1.25, 1, 0.75, 0.5).$$

A total of $B = 100$ Monte Carlo (MC) samples are produced in both the simulation and the sensitivity study. Results are reported in the next Sections.

5.2 Simulation experiment with fixed level of contamination

From the DGP outlined in Section 5.1, $N = 500$ units are generated and their group membership retained for constructing the training set; while $M = 5000$ unlabeled observations compose the test set. Subsequently, label noise is simulated by wrongly assigning 20 units coming from the fourth group to the third class. In addition, 5 uniformly distributed outliers, having

- squared Mahalanobis distances $d(\cdot; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ greater than $\chi_{3,0.975}^2 \forall g \in \{1, 2, 3, 4\}$ for the relevant variables
- $d(\cdot; \boldsymbol{\mu}_g^{[1,3]} \mathbf{B}, \mathbf{I}_4)$ greater than $\chi_{4,0.975}^2 \forall g \in \{1, 2, 3, 4\}$ for the redundant variables
- $d(\cdot; \boldsymbol{\eta}, \boldsymbol{\delta})$ greater than $\chi_{9,0.975}^2$ for the irrelevant variables

are appended to the training set, with randomly assigned labels. This contamination produce, in each MC replication, a total of 25 adulterated units, that account for slightly less than 5% of the entire learning set. We validate the performance of our novel methods in correctly retrieving the relevant variables, compared to non-robust procedures. Particularly, the comparison is carried out considering the following methods:

- TBIC: robust stepwise greedy-forward approach via TBIC (Section 4.1)
- ML subset: maximum likelihood subset selector approach (Section 4.2), with subset size of relevant variables p equal to 3, 6 and 9
- SRUW: stepwise greedy-forward approach via BIC (Maugis et al., 2011)
- SelvarMix: variable selection in model-based discriminant analysis with a regularization approach (Celeux et al., 2019).

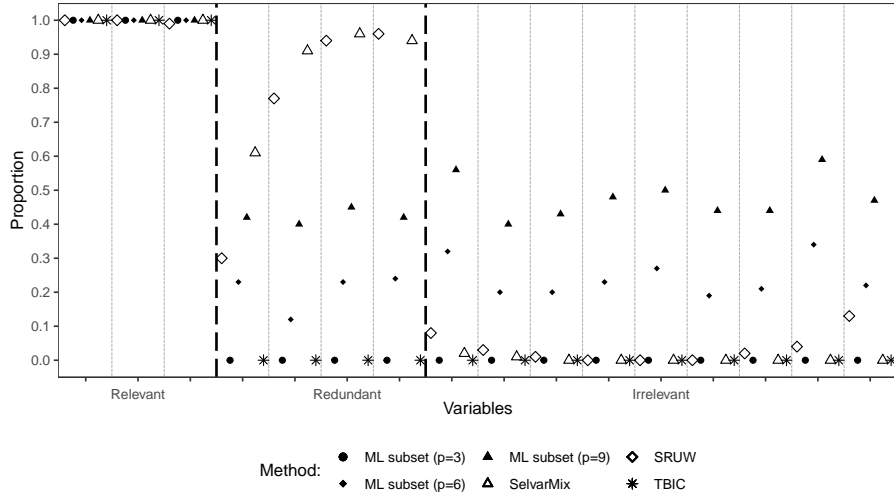


Figure 3: Proportion of times a variable has been selected as relevant, out of $B = 100$ MC repetition of the simulated experiment, for different variable selection methods.

Furthermore, once the important variables have been identified, the associated classifier (i.e., REDDA for the robust variable selection criteria and EDDA for the non-robust ones) is trained on the reduced set of predictors and the classification accuracy is computed on the test set. A labeled trimming level γ equal to 0.05 was kept fixed during the experiment. Lastly, for providing benchmark values on the relevance of feature selection, both EDDA and REDDA classifiers are also fitted on the original set with $P = 16$ variables.

Figure 3 displays the proportion of times a variable has been selected as relevant by the different methods in the $B = 100$ repetitions of the simulated experiment. As it is clearly visible from the plot, the first three features are selected by all the procedures in almost every iteration of the simulation study. Generally, therefore, the contamination introduced in the training set does not cause any systematic exclusion of the true discriminative variables from the relevant subset, also for the non-robust methods. Nonetheless, outliers and label noise lead SRUW and SelvarMix to severely overestimate the number of retained features. Redundant and irrelevant variables are often included in the selection, as demonstrated by the hollow triangles and diamonds in Figure 3. The robust stepwise approach via TBIC, instead, does not seem to suffer from this unfavorable behavior: it correctly identifies the first three relevant variables in every single simulation. As already pointed out in Section 4.3, the main drawback of the maximum likelihood subset selector approach is given by the need of pre-specifying the subset size p . When $p = 3$, i.e., the true number of discriminating variables, the algorithm always correctly selects the relevant ones. Clearly, when p is set higher than three, some irrelevant and/or redundant

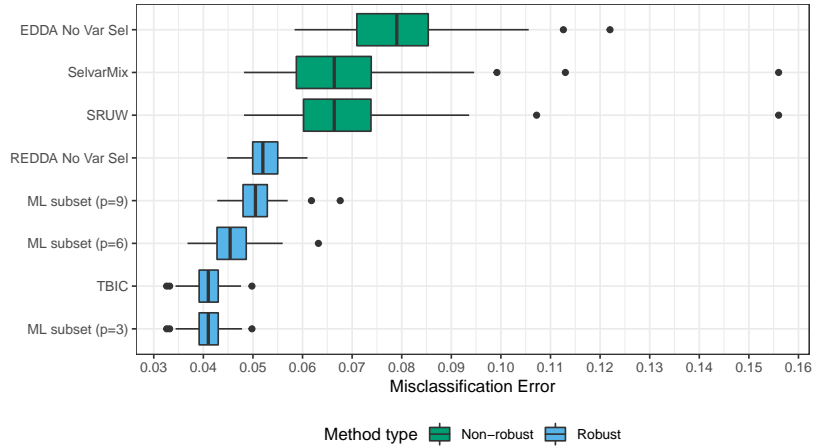


Figure 4: Boxplots of the misclassification error, out of $B = 100$ MC repetition of the simulated experiment, for the $M = 5000$ test data, varying variable selection and model-based classification methods.

features will be necessarily included in the retained set. However, letting p to be greater than the true relevant predictors does not seem to severely affect the predictive power of the robust classification rule. As it can be seen from the results reported in Table 1 and in Figure 4, the misclassification errors are only slightly influenced by the choice of p in the ML subset selector, and are always on average lower than non-robust procedures. As expected, the best prediction accuracy is obtained when $p = 3$, result that entirely agrees with the one obtained by the forward selection algorithm via TBIC, as the very same variables are selected for each simulation and, subsequently, the REDDA classifier is fitted on the retained subset. Interestingly, the EDDA classifier

Table 1: Average misclassification error, out of $B = 100$ MC repetition of the simulated experiment, for the $M = 5000$ test data, varying variable selection and model-based classification methods. Standard deviations reported in parentheses.

Method	Misclassification Error	Method	Misclassification Error
ML subset ($p=3$)	0.0411 (0.003)	REDDA No Var Sel	0.0525 (0.003)
ML subset ($p=6$)	0.0457 (0.0045)	SRUW	0.0686 (0.0045)
ML subset ($p=9$)	0.0506 (0.004)	SelvarMix	0.0684 (0.004)
TBIC	0.0411 (0.003)	EDDA No Var Sel	0.0795 (0.003)

coupled with (non-robust) variable selection via either SelvarMix or SRUW shows on average higher misclassification error than REDDA learned on the entire set of features. That is, the harmful effect of adulterated observations is increased by the presence of noisy variables, also shown by the poor performance of EDDA with no feature selection. The present simulation study highlights how a very small proportion of attribute and class noise may somewhat spoil a wrapper procedure, driving the algorithm to include many more features than the truly relevant ones. That is, when adulterated units are not properly dealt with, both feature identification and classification may provide inappropriate results, with bias in the former propagating to badly affect the derived classifier even further.

5.3 Sensitivity study varying contamination and trimming levels

A sensitivity study is built upon the same DGP outlined in the previous Sections, encompassing 9 different scenarios varying the true proportions of attribute and class noise in the training set. The actual number of mislabeled and outlying units, and resultant contamination rate are reported in Table 2. The aim of this experiment is to numerically investigate the effect that the misspecification of γ , i.e., to overestimate or to underestimate the true contamination level, produces in the novel variable selection procedures. The results obtained in Section 5.2

Table 2: Number of outliers, label noise and resultant true contamination rate for the 9 scenarios encompassing the sensitivity study.

# Outliers	0	0	0	30	30	30	50	50	50
# Label Noise	0	30	50	0	30	50	0	30	50
Contamination	0	0.06	0.10	0.057	0.113	0.151	0.091	0.145	0.182

display a general tendency in overstating the important subset magnitude for non-robust solutions. Motivated by this fact, we are most interested in evaluating the ability of our methods in solely retrieving the 3 relevant features. To this extent, we employ the “variable selection precision” metric as a measure of performance, defined to be the proportion of features selected by a methodology truly belonging to the relevant subset, out of all the retained ones. For each scenario reported in Table 2, we fit both the robust stepwise via TBIC and the ML subset selector (with p fixed and equal to 3) approaches considering impartial trimming levels γ respectively equal to 0, 0.05 and 0.15. In addition, we also evaluate the precision of the so-called “oracle” solutions, obtained by setting γ equal to the true contamination rate. Sensitivity study results are summarized in Figure 5, in which several interesting patterns emerge.

First off, it is immediately noticed that the overall precision is much more badly affected by outliers addition rather than by mislabeled samples. Such a behavior is explained by recognizing that label noise does not directly alter the feature space: classes separation gets indeed more uncertain, yet the most

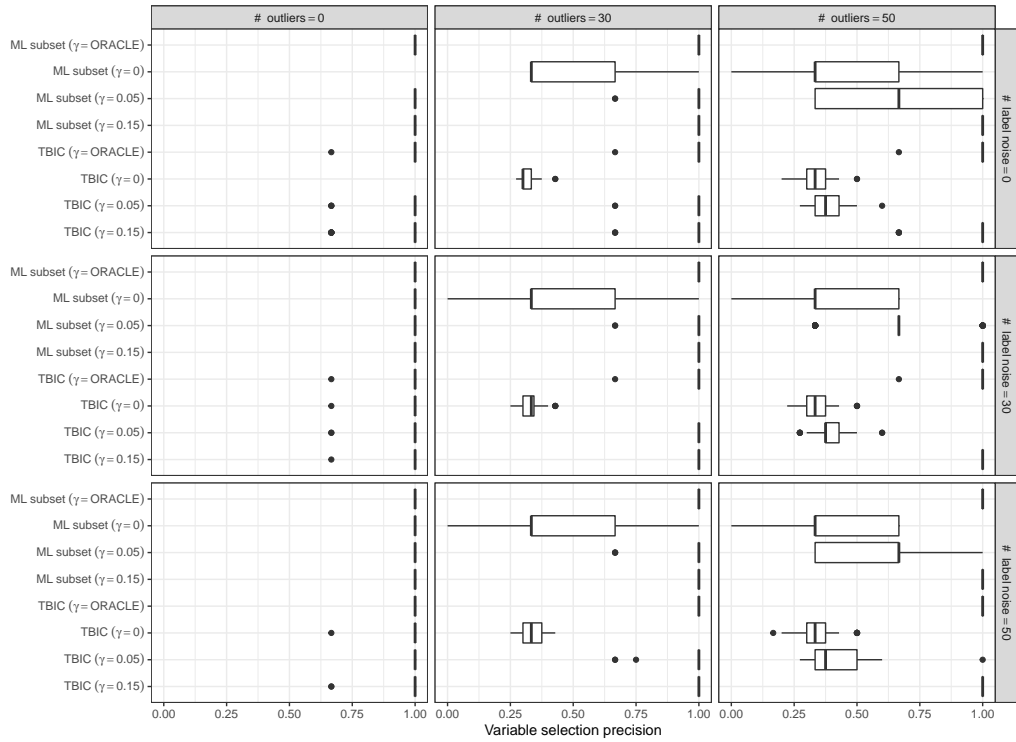


Figure 5: Boxplots of variable selection precision, out of $B = 100$ MC repetition of the sensitivity study, varying contamination and trimming levels. The boxplots of TBIC ($\gamma = 0$) and ML subset ($\gamma = 0$) for the uncontaminated scenario (top left corner) are not displayed since they agree with the oracle solution. For the ML subset selector approach, the subset size of relevant variables p is set equal to 3.

discriminative variables are still identified even when γ is underestimated. The same does not happen when it comes to attribute noise; as soon as the number of outliers exceeds the trimming level the obtained precision deteriorates for both methodologies. Notwithstanding, solutions based on high values of γ (i.e., $\gamma = 0.15$) adequately safeguard the procedure performance, even in the most extreme scenario (for which the true contamination level is equal to 0.182).

Secondly, we observe that, conditioning on the same γ value, the approach based on stepwise TBIC generally shows worse performances with respect to the ML subset selector method. Needless to say, the fact of a-priori setting the number of retained variables p equal to the true relevant subset has a major positive impact for the latter method. In details, the stepwise TBIC procedure tends to include unnecessary features when γ is underestimated, with consequent loss in variable selection precision. On the other hand, the ML subset selector is forced to select only the $p = 3$ most relevant ones. In spite of that, for the scenarios in which the outliers proportion exceeds the trimming level we notice a reduction in terms of estimated precision: the method fails in distinguishing between redundant and relevant features.

Lastly, and perhaps most importantly, the results highlight that cautiously overestimating the contamination level does not impact the ability of our methods in retrieving the relevant subset; the same cannot be said when γ is set lower than needed. The “trimming more is better than trimming less” principle is well known for robust methods based on hard-trimming, it seems however particularly true for our variable selection procedures, for which no corresponding drawback to the efficiency loss in parameter estimation has been identified in our synthetic experiments. Clearly, further considerations are needed to formally assess this promising property, and, to this extent, a separate line of research is currently being pursued.

All in all even though, as already pointed out in Section 4.4, properly choosing γ still remains a critical step in all robust procedures based on impartial trimming, this sensitivity study underlines how variable selection seems not to be badly affected by an overestimation of the contamination level. Therefore, replacing standard methods with robust solutions seems paramount whenever it is believed the considered dataset may contain some noisy units, especially in high dimensional settings.

6 Application to MIR spectra: starches discrimination

Chemometrics is a natural field of application for high-dimensional statistics, as data recorded from chemical systems are complex in nature and generally limited in terms of sample size. In particular, variable selection methods are notably appealing for observations recorded by spectroscopic instruments: for virtually continuous spectra the information contained in adjacent features is often correlated, and thus the determination of a relevant subset of wavelengths

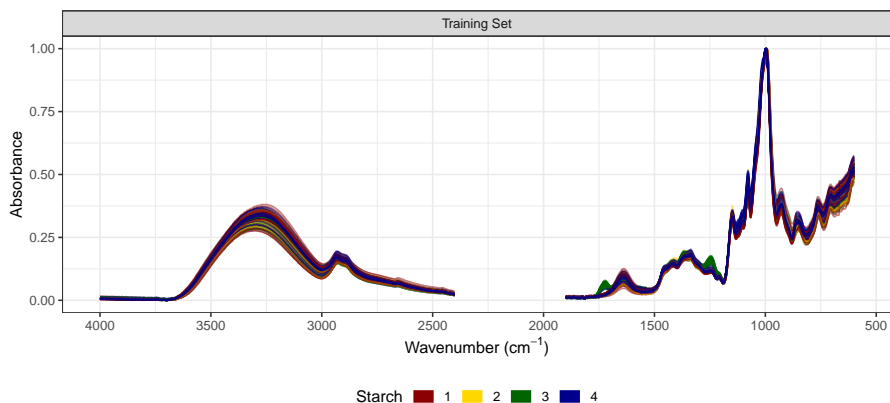


Figure 6: Mid infrared spectra of starches of four different classes, training set.

is desirable, prior to perform any subsequent analysis (Brown, 1992; Brenchley et al., 1997). Furthermore, data reduction simplifies results interpretation, making future measurements simpler and cheaper (Indahl and Næs, 2004).

Spectroscopic data are recorded during a controlled experiment, and the quality of both measurements and analysed substances is, in most cases, reliable. Nevertheless, calibration errors may appear during spectra collection, and, moreover, for some delicate applications such as food authenticity, the raw material itself may be spoiled and/or adulterated (Reid et al., 2006). In this context, therefore, variable selection methods that not only robustly identify relevant wavelengths, but also recognize outliers and possibly fraudulent samples may be particularly valuable to chemometricians. Motivated by a Mid-infrared (MIR) dataset of the chemometrics challenge organized during the ‘Chimiométrie 2005’ conference, the methodologies introduced in Section 4 are employed for performing high-dimensional classification and outlier detection.

6.1 Data

The considered datasets, described in Fernández Pierna et al. (2005); Fernández Pierna and Dardenne (2007), include respectively $N = 215$ (training set) and $M = 43$ (test set) MIR spectra of starches of four different classes, taken on a Perkin-Elmer Spectrum 2000 FTIR spectrometer (Perkin Elmer Corporation, Norwalk, CT, USA) between 4000 and 600 cm^{-1} at 1 cm^{-1} data interval. The range between 2402 and 1901 cm^{-1} was removed from the spectra, so that a total of $P = 2901$ absorbance measurements were then retained for the analysis. A subset of the learning observations is displayed in Figure 6. In order to create an extra difficulty to be tackled by the participants during the competition, four outliers were included in the test set:

- *Sample 2*: a shifted version of unit 1, obtained by removing its first six data points and appending six new variables at the end of the spectrum;

- *Sample 4*: a noisy version of unit 2, by generating Gaussian white noise and adding it to the absorbance values of the sample;
- *Sample 43*: a modified version of unit 39, obtained by manually changing a data point on the spectrum (wavelength 2456) to simulate a spike;
- *Sample 20*: a modified version of unit 17, by adding a slope to its original spectrum.

Therefore, the discrimination challenge held during ‘Chimiométrie 2005’ consisted in learning a classification rule from the training set to predict the labels of the test units, whilst also performing adulteration detection on the latter. In our experiment, we additionally include label noise by wrongly assigning the last four units of the third group of starches to the fourth one: this accounts for less than 2% of the entire training set. Classification results are reported in the next Section.

6.2 Results

The discriminating problem described in the previous Section cannot be solved by directly applying model-based classifiers, since $N \ll P$. To overcome this issue, we make use of the robust wrapper variable selection methods introduced in this article: such approaches provide a natural solution for dealing with contaminated high-dimensional data, and, as we will see, they can be further used to identify the noisy units in the test set. We firstly run the stepwise greedy-forward approach via TBIC (Section 4.1) with $\gamma = 0.05$: the procedure, out of $P = 2901$, selects a total of only six relevant wavelengths: 1728 cm^{-1} , 1682 cm^{-1} , 1555 cm^{-1} , 1502 cm^{-1} , 997 cm^{-1} and 995 cm^{-1} . Figure 7 displays the generalized pairs plot for the selected variables. Motivated by the TBIC output and by the results presented in the Simulation Study, we decided to retain a slightly higher number of relevant variables in the ML subset selector, setting the value of p to be equal to 9. In doing so, the ML subset selector estimates the relevant subset F to be comprised of the following wavelengths: 998 cm^{-1} , 1089 cm^{-1} , 1223 cm^{-1} , 1255 cm^{-1} , 1311 cm^{-1} , 1565 cm^{-1} , 1647 cm^{-1} , 1711 cm^{-1} and 1754 cm^{-1} . A generalized pairs plot (Emerson et al., 2013) of such subset is reported in Figure 8. Interestingly, the two approaches select entirely different wavelengths as the most discriminative ones. Careful investigation of this behavior shows high correlation between the variables selected by the two methodologies, while the correlation reported by features within the same subset is much lower. Clearly, in dealing with real datasets the separation between relevant, irrelevant and redundant variables is much less apparent. Particularly for spectroscopic data, highly correlated wavelengths often result in comparable discriminating power, with no natural preference in terms of relevance. Nevertheless, it is worth noting that both methods chose wavelengths from the right-hand side part of the spectrum, as it seems to delineate the highest separation between the different starches, also by visual inspection of Figure 6.

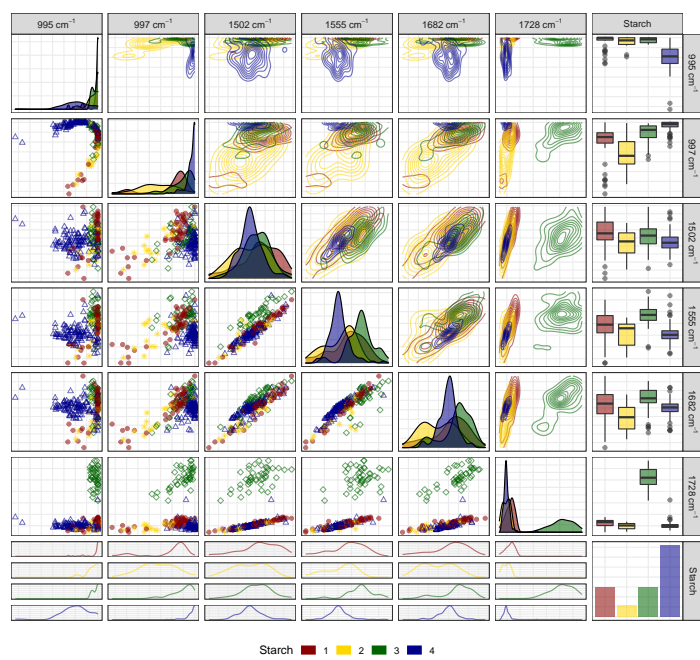


Figure 7: Generalized pairs plot of the relevant variables selected by the stepwise greedy-forward approach via TBIC. Starches dataset, training samples.

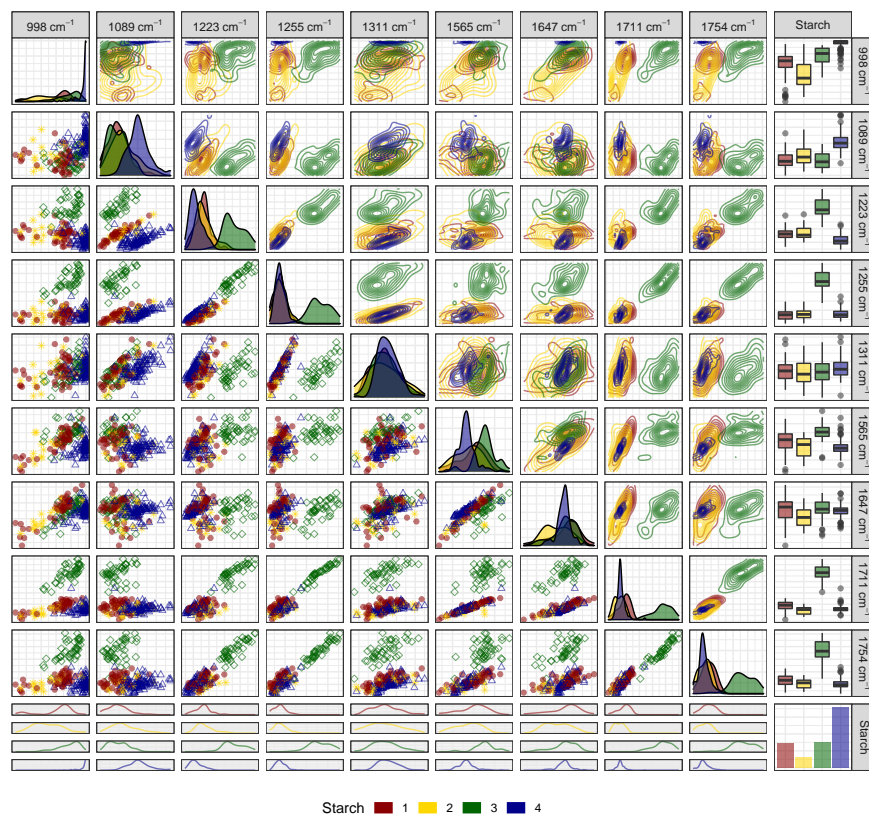


Figure 8: Generalized pairs plot of the relevant variables selected by ML subset selector with $p = 9$. Starches dataset, training samples.

A REDDA model with $\gamma = 0.05$ is employed to predict the class for the test samples, using as predictors the variables retained by the TBIC and ML subset selector, respectively. In both cases, units that present class noise in the training set were correctly identified as such and not accounted for in the estimation procedure. In addition, 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 (Fernández Pierna et al., 2005; Fernández Pierna and Dardenne, 2007). Lastly, we replicate the second best solution proposed by one of the 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 four competing methods, considering test sets without modified units, is reported in Table 3:

Table 3: Number of correctly predicted test samples and associated misclassification error for different methods. The test set without outliers has a total sample size of $M = 39$.

	REDDA (TBIC)	REDDA (ML subset)	SVM radial kernel	ROC+PLS+SVM
#correctly predicted	32	34	31	31
Misclassification error	0.179	0.128	0.205	0.205

the robust model-based classifiers show better results than the other solutions. The performance of the kernel and ensemble methods are negatively impacted by the presence of the 4 mislabelled units in the training set: compare results in Table 3 with the ones reported in Table 1 of Fernández Pierna and Dardenne (2007), wherein the classifiers were trained on an uncontaminated learning set. The relevant subsets retained by both robust variable selection methods lead to similar results in terms of classification accuracy, with a slight better performance when REDDA is fitted on the features identified by the ML subset selector approach. As already pointed out in Fernández Pierna and Dardenne (2007), the main source of error is due to the difficulties in separating classes 1 and 2, as it is evident also in Figures 7 and 8.

We mentioned at the beginning of the Section that the REDDA method can be effectively employed in performing outlier detection in the test set. Particularly, given the probabilistic assumptions that underlie the methodology, for each test unit \mathbf{y}_m , $m = 1, \dots, M$, we can compute its estimated marginal density as follows:

$$\hat{p}(\mathbf{y}_{m,\hat{F}}; \hat{\tau}, \hat{\boldsymbol{\mu}}_{\hat{F}}, \hat{\boldsymbol{\Sigma}}_{\hat{F}}) = \sum_{g=1}^G \hat{\tau}_g \phi(\mathbf{y}_{m,\hat{F}}; \hat{\boldsymbol{\mu}}_{g,\hat{F}}, \hat{\boldsymbol{\Sigma}}_{g,\hat{F}}) \quad (22)$$

where \hat{F} denotes either the relevant variables identified by the stepwise approach with TBIC or by the ML subset selector, with parameters robustly estimated via the REDDA model on the retained features. For both variable selection

approaches, the 3 observations \mathbf{y}_m with lowest value of (22) are units 2, 4 and 20; all of them were manually modified, as described in Section 6.1. The only neglected outlier is unit 43: it was contaminated on a single wavelength that was not identified as relevant by the variable selection methods. Nonetheless, by using an impartial trimming approach, we are effectively able to identify 3 out of 4 adulterated units.

In this Section, we have shown that the proposed noise-resistant variable selection approaches, coupled with robust discriminant analysis, can be effectively employed in performing high-dimensional classification in an adulterated framework. Even though being notably noise tolerant, powerful classifiers such as Support Vector Machine provide lower classification accuracy when a small percentage of class noise is present in the training set. In addition, after parameters have been robustly estimated, our proposal can be used to recognize possible adulterated units in the test set. All in all, an automatic methodology that performs robust feature detection, parameter estimation and outlier identification may become beneficial in chemometrics, easing both pre and post processing steps of complex spectroscopic analyses.

7 Concluding Remarks

In the present manuscript we have introduced two wrapper variable selection methods, resistant to outliers and label noise. We have shown that by means of these approaches we can effectively perform high-dimensional discrimination in an adulterated scenario. The first wrapper method embeds a robust model-based classifier within a greedy-forward algorithm, validating stepwise inclusion and exclusion of variables from the relevant subset via a robust information criterion. Some theoretical justifications that corroborates the procedure are also discussed. The second wrapper method resorts to the theory of maximum likelihood and irrelevance, defining an objective function in which the subset of relevant variables is regarded as a parameter to be estimated. A dedicated algorithm for MLE within a Gaussian family of patterned models has been developed, and practical implementation issues have been considered. Further, pros and cons of the two novel procedures have been discussed. The robust stepwise approach via TBIC enjoys the automatic identification of the relevant subset size, and it has displayed less variability in terms of selection precision. On the other hand, the ML subset selector is computationally faster and can be specifically useful in situations for which the number of important variables is known in advance, even though being this occasionally the case in applications. A simulation study has been developed for assessing the effectiveness of our proposals in recovering the true discriminative features in a contaminated scenario, comparing their performances against well-known variable selection criteria. The novel methods have then been successfully applied in solving a high-dimensional classification problem of contaminated spectroscopic data. High discriminating power has been exhibited by the final models, whence the identification of the wrongly labeled and/or adulterated observations is derived

as a by-product of the estimation procedures.

An open point for further research regards the extension of the fully supervised framework outlined here to the adaptive one, where unobserved classes in the test set need also to be discovered, embedding the resulting semi-supervised procedure within a robust variable selection approach. In addition, careful investigation will be devoted to the development of a methodology that automatically assesses the contamination rate present in a sample, as the a-priori specification of the trimming level still remains an open issue in this field, particularly delicate for high-dimensional data.

Acknowledgments

The authors are grateful to Prof. Dr. Gunter Ritter for the stimulating discussions and suggestions on how to transpose the ML subset selector approach, originally developed for clustering, to the classification framework. Thanks are due to Professor Ludovic Duponchel for providing relevant context on how the novel methodologies may be favorably employed, as well as for supplying the dataset. The authors also thank the editor and the two anonymous referees: their valuable comments greatly improved the quality of the paper. Brendan Murphy's work is supported by Science Foundation Ireland grants (SFI/12/RC/2289_P2 and 16/RC/3835). Andrea Cappozzo and Francesca Greselin's work is supported by Milano-Bicocca University Fund for Scientific Research, 2019-ATE-0076.

A Further aspects for the robust stepwise greedy-forward approach via TBIC

In this Section we retrieve the ML estimates for the grouping and no grouping structures in the robust stepwise greedy-forward approach (Section 4.1), by means of the spurious outliers model specification.

Grouping Model

The log-likelihood function of the spurious outliers model under the grouping structure is:

$$\begin{aligned} \ell(\mathcal{N}, \boldsymbol{\tau}^{cp}, \boldsymbol{\mu}^{cp}, \boldsymbol{\Sigma}^{cp}) &= \sum_{n=1}^N q_n \sum_{g=1}^G l_{ng} \log(\tau_g^{cp} \phi(\mathbf{x}_n^c, x_n^p; \boldsymbol{\mu}_g^{cp}, \boldsymbol{\Sigma}_g^{cp})) + \\ &+ \sum_{n=1}^N (1 - q_n) \log w(\mathbf{x}_n^c, x_n^p, \mathbf{l}_n; \boldsymbol{\psi}_n) \end{aligned} \quad (23)$$

to be maximized with respect to $\{\mathcal{N}, \boldsymbol{\tau}^{cp}, \boldsymbol{\mu}^{cp}, \boldsymbol{\Sigma}^{cp}\}$. The problem then reads:

$$\begin{aligned} \max_{\mathcal{N} \in \mathcal{D}(N)} \left[\max_{\boldsymbol{\tau}^{cp}, \boldsymbol{\mu}^{cp}, \boldsymbol{\Sigma}^{cp}} \sum_{n=1}^N q_n \sum_{g=1}^G l_{ng} \log (\tau_g^{cp} \phi(\mathbf{x}_n^c, x_n^p; \boldsymbol{\mu}_g^{cp}, \boldsymbol{\Sigma}_g^{cp})) + \right. \\ \left. + \max_{\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_N} \sum_{n=1}^N (1 - q_n) \log w(\mathbf{x}_n^c, x_n^p, \mathbf{l}_n; \boldsymbol{\psi}_n) \right]. \end{aligned} \quad (24)$$

By property (15), any configuration that maximizes the first addend in (24) also maximizes the second one. For a fixed partition $\mathcal{N} \in \mathcal{D}(N)$, the MLE for the first quantity are given by:

$$\begin{aligned} \hat{\tau}_g^{cp} &= \frac{\sum_{n=1}^N q_n l_{ng}}{[N(1 - \gamma)]} \quad g = 1, \dots, G \\ \hat{\boldsymbol{\mu}}_g^{cp} &= \frac{\sum_{n=1}^N q_n l_{ng}(\mathbf{x}_n^c, x_n^p)}{\sum_{n=1}^N q_n l_{ng}} \quad g = 1, \dots, G. \end{aligned}$$

Estimation of $\boldsymbol{\Sigma}_g^{cp}$ depends on the considered patterned model, details are given in Bensmail and Celeux (1996). Operatively, the final estimates are obtained via a REDDA model fitted on \mathbf{x}_n^c, x_n^p , see Section 3.

No grouping Model

The log-likelihood function of the spurious outliers model under the no grouping structure is:

$$\begin{aligned} \ell(\mathcal{D}, \boldsymbol{\tau}^c, \boldsymbol{\mu}^c, \boldsymbol{\Sigma}^c, \alpha, \boldsymbol{\beta}, \sigma^2) &= \sum_{n=1}^N q_n \sum_{g=1}^G l_{ng} \log [\tau_g^c \phi(\mathbf{x}_n^c; \boldsymbol{\mu}_g^c, \boldsymbol{\Sigma}_g^c)] + \\ &+ \sum_{n=1}^N q_n \log [\phi(x_n^p; \alpha + \boldsymbol{\beta}' \mathbf{x}_n^r, \sigma^2)] + \\ &+ \sum_{n=1}^N (1 - q_n) \log w(\mathbf{x}_n^c, x_n^p; \boldsymbol{\psi}_n) \end{aligned} \quad (25)$$

to be maximized with respect to $\{\mathcal{N}, \boldsymbol{\tau}^c, \boldsymbol{\mu}^c, \boldsymbol{\Sigma}^c, \alpha, \boldsymbol{\beta}, \sigma^2\}$. The problem then reads:

$$\begin{aligned} \max_{\mathcal{N} \in \mathcal{D}(N)} \left[\max_{\boldsymbol{\tau}^c, \boldsymbol{\mu}^c, \boldsymbol{\Sigma}^c} \sum_{n=1}^N q_n \sum_{g=1}^G l_{ng} \log [\tau_g^c \phi(\mathbf{x}_n^c; \boldsymbol{\mu}_g^c, \boldsymbol{\Sigma}_g^c)] + \right. \\ \left. + \max_{\alpha, \boldsymbol{\beta}, \sigma^2} \sum_{n=1}^N q_n \log [\phi(x_n^p; \alpha + \boldsymbol{\beta}' \mathbf{x}_n^r, \sigma^2)] + \right. \\ \left. + \max_{\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_N} \sum_{n=1}^N (1 - q_n) \log w(\mathbf{x}_n^c, x_n^p, \mathbf{l}_n; \boldsymbol{\psi}_n) \right]. \end{aligned} \quad (26)$$

By property (15), any configuration that maximizes the sum of the first and second term in (26) also maximizes the third one. For a fixed partition $\mathcal{N} \in \mathcal{D}(N)$, the first two quantities can be separately maximized, leading to the following MLE

$$\begin{aligned}\hat{\tau}_g^c &= \frac{\sum_{n=1}^N q_n l_{ng}}{\lceil N(1-\gamma) \rceil} \quad g = 1, \dots, G \\ \hat{\boldsymbol{\mu}}_g^c &= \frac{\sum_{n=1}^N q_n l_{ng} \mathbf{x}_n^c}{\sum_{n=1}^N q_n l_{ng}} \quad g = 1, \dots, G.\end{aligned}$$

for the former term, where as usual $\hat{\Sigma}_g^c$ depends on the considered patterned model. ML estimates for the regression coefficients are obtained solving the following minimization problem:

$$\min_{\alpha, \boldsymbol{\beta}} \sum_{n=1}^N q_n (x_n^p - \alpha - \boldsymbol{\beta}' \mathbf{x}_n^r)^2 \quad (27)$$

which is very similar to the least trimmed squares method (Rousseeuw, 1984). Lastly, the variance is estimated as follows:

$$\hat{\sigma}^2 = \frac{1}{\lceil N(1-\gamma) \rceil} \sum_{n=1}^N q_n (x_n^p - \hat{\alpha} - \hat{\boldsymbol{\beta}}' \mathbf{x}_n^r).$$

Operatively, the MLE for (25) are obtained combining a REDDA model on \mathbf{x}_n^c with a robust linear regression of x_n^p on \mathbf{x}_n^r . The discriminating function in (12) is used to determine the subset of untrimmed units on which to compute the estimates defined above, iterating the algorithm until the same observations are discarded in two consecutive steps. Lastly, at each iteration, similarly to what performed in `clustvsel` (Scrucca and Raftery, 2018), the subset of variables \mathbf{x}_n^r is determined with the `bicreg` function in the `BMA` R package (Raftery et al., 2018).

B Further aspects for the ML subset selector approach

This final Section discusses the computational details of the algorithm used for fitting the ML subset selector, whose main steps are reported in Section 4.2. For achieving flexibility, parsimony and computational speed, the family of patterned models based on the eigenvalue decomposition in (3) of Bensmail and Celeux (1996) is considered. Particularly, we adopt the three-letter identifier used in the `mclust` software for naming the models, where the volume, shape and orientation can be either equal (E) or different (V) across groups, with full (**E, **V), diagonal (**I) or spherical (*II) components: we refer to Scrucca et al. (2016) for the complete details. Let us further introduce the following notations: for a $d \times d$ matrix \mathbf{A} , $\text{diag}(\mathbf{A})$ denotes the $d \times d$ diagonal matrix whose diagonal entries are the same of the matrix \mathbf{A} . Lastly, $\mathbf{A}(i, j)$ denotes the scalar entry at the i th row and j th column of the matrix \mathbf{A} .

Computational details on the M-step

As previously mentioned, we refer the reader to Bensmail and Celeux (1996) for a complete treatment on the estimation of Σ_g , $g = 1, \dots, G$ under the 14 covariance structures. Conditioning on the chosen model, the estimation of the pooled covariance matrix Σ has the following form:

- Ellipsoidal:

$$\hat{\Sigma}_{ell} = \frac{1}{\lceil N(1-\gamma) \rceil} \sum_{n=1}^N \zeta(\mathbf{x}_n) \left[(\mathbf{x}_n - \hat{\boldsymbol{\mu}})(\mathbf{x}_n - \hat{\boldsymbol{\mu}})' \right]$$

for EEE, VEE, EVE, EEV, VVE, VEV, EVV and VVV models

- Diagonal:

$$\hat{\Sigma}_{diag} = \text{diag}(\hat{\Sigma}_{ell})$$

for EEI, VEI, EVI, VVI models.

- Spherical:

$$\hat{\Sigma} = \frac{1}{P} \sum_{d=1}^P \hat{\Sigma}_{diag}(d, d) \mathbf{I}_P$$

for EII, VII models

Computational details on the S-step

The S-step involves a discrete structure optimization, where we seek to determine the set of p variables that minimizes (21). Solving the problem by exhaustive enumeration is feasible only when $\binom{P}{p}$ is not too large, sadly it is rarely the case in a high-dimensional setting. Thus, the considered implementation relies on a stochastic algorithm for fixed-size subset selection, by means of the `kofnGA` R package (Wolters, 2015). Nonetheless, for specific patterned structures, simpler form of the objective function may be derived: see the following sections.

EEE model

For the homoscedastic model (EEE), (21) simplifies as follows:

$$h(F) = \log \det \hat{\Sigma}_{EEE,F} - \log \det \hat{\Sigma}_{ell,F} \quad (28)$$

where

$$\hat{\Sigma}_{EEE,F} = \frac{1}{\lceil N(1-\gamma) \rceil} \sum_{g=1}^G \hat{n}_g \sum_{n=1}^N \zeta(\mathbf{x}_n) \left[(\mathbf{x}_{n,F} - \hat{\boldsymbol{\mu}}_{g,F})(\mathbf{x}_{n,F} - \hat{\boldsymbol{\mu}}_{g,F})' \right]$$

and

$$\hat{n}_g = \frac{\sum_{n=1}^N \zeta(\mathbf{x}_n) l_{ng}}{\lceil N(1-\gamma) \rceil}.$$

It is nevertheless computationally efficient to derive $\hat{\Sigma}_{EEE}$ for the full dimension P at once and to extract the sub-matrix $\hat{\Sigma}_{EEE,F}$ when needed.

VVI model

For the heteroscedastic diagonal model (VVI), (21) simplifies to:

$$h(F) = \sum_{k \in F} \sum_{g=1}^G \hat{\tau}_g \log \frac{\hat{\Sigma}_g(k, k)}{\hat{\Sigma}_{diag}(k, k)} \quad (29)$$

for which \hat{F} is the set of the indices k with the p smallest sums $\sum_{g=1}^G \hat{\tau}_g \log \frac{\hat{\Sigma}_g(k, k)}{\hat{\Sigma}_{diag}(k, k)}$.

EEl model

For the homoscedastic diagonal model (EEl), (21) reads:

$$h(F) = \sum_{k \in F} \log \frac{\hat{\Sigma}_{EEI}(k, k)}{\hat{\Sigma}_{diag}(k, k)} \quad (30)$$

with

$$\hat{\Sigma}_{EEI} = \frac{1}{[N(1-\gamma)]} \sum_{g=1}^G \hat{n}_g \text{diag} \left(\sum_{n=1}^N \zeta(\mathbf{x}_n) \left[(\mathbf{x}_{n,F} - \hat{\boldsymbol{\mu}}_{g,F})(\mathbf{x}_{n,F} - \hat{\boldsymbol{\mu}}_{g,F})' \right] \right).$$

In this case, \hat{F} is the set of the indices k with p smallest quotients $\frac{\hat{\Sigma}_{EEI}(k, k)}{\hat{\Sigma}_{diag}(k, k)}$.

Computational details on the T-step

When the full dimension P is large, it may occur that $\hat{\Sigma}_{\hat{E}|\hat{F}}$ is not of full rank. In this case, it is still possible to estimate a singular normal distribution on a subspace of the set \hat{E} of irrelevant variables. The associated density will then be:

$$\frac{(2\pi)^{-k/2}}{\left(\prod_{k=1}^K \omega_k\right)^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_{n,\hat{E}} - \hat{\mathbf{G}}_{\hat{E}|\hat{F}} \mathbf{x}_{n,\hat{F}} - \hat{\boldsymbol{\mu}}_{\hat{E}|\hat{F}})' \hat{\Sigma}_{\hat{E}|\hat{F}}^- (\mathbf{x}_{n,\hat{E}} - \hat{\mathbf{G}}_{\hat{E}|\hat{F}} \mathbf{x}_{n,\hat{F}} - \hat{\boldsymbol{\mu}}_{\hat{E}|\hat{F}}) \right\} \quad (31)$$

where $\hat{\Sigma}_{\hat{E}|\hat{F}}^-$ is the g-inverse of $\hat{\Sigma}_{\hat{E}|\hat{F}}$ and $\omega_1, \dots, \omega_K$ are the non-zero eigenvalues of $\hat{\Sigma}_{\hat{E}|\hat{F}}$.

Models comparison

As a final remark, we mention the possibility of developing a procedure for automatically choosing the best model within the 14 parsimonious structures in the ML subset selector approach. One could rely on a BIC-like criterion

(Schwarz, 1978), penalizing twice the final maximized trimmed log-likelihood by the number of estimated parameters and untrimmed observations, retaining the model that presents the highest value. However, this would rapidly increase the computational time needed for performing the analysis. For this reason, in both the simulation study and in the application, a VVV model only was considered when fitting the ML subset selector approach.

References

- Andrews JL, McNicholas PD (2014) Variable Selection for Clustering and Classification. *Journal of Classification* 31(2):136–153
- Banfield JD, Raftery AE (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics* 49(3):803
- Bellman R (1957) *Dynamic Programming*. Rand Corporation research study, Princeton University Press
- Bensmail H, Celeux G (1996) Regularized Gaussian Discriminant Analysis through Eigenvalue Decomposition. *Journal of the American Statistical Association* 91(436):1743–1748
- Blum AL, Langley P (1997) Selection of relevant features and examples in machine learning. *Artificial Intelligence* 97(1-2):245–271
- Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M (2020) Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis* 143:106839
- Bouveyron C, Brunet-Saumard C (2014) Model-based clustering of high-dimensional data: A review. *Computational Statistics and Data Analysis* 71:52–78
- Bouveyron C, Celeux G, Murphy TB, Raftery AE (2019) *Model-Based Clustering and Classification for Data Science*, vol 50. Cambridge University Press
- Brenchley JM, Hörchner U, Kalivas JH (1997) Wavelength Selection Characterization for NIR Spectra. *Applied Spectroscopy* 51(5):689–699
- Brown PJ (1992) Wavelength selection in multicomponent near-infrared calibration. *Journal of Chemometrics* 6(3):151–161
- Cappozzo A, Greselin F, Murphy TB (2020) A robust approach to model-based classification based on trimming and constraints. *Advances in Data Analysis and Classification* 14(2):327–354
- Celeux G, Govaert G (1995) Gaussian parsimonious clustering models. *Pattern Recognition* 28(5):781–793

- Celeux G, Maugis-Rabusseau C, Sedki M (2019) Variable selection in model-based clustering and discriminant analysis with a regularization approach. *Advances in Data Analysis and Classification* 13(1):259–278
- Cerioli A, Riani M, Atkinson AC, Corbellini A (2018) The power of monitoring: how to make the most of a contaminated multivariate sample. *Statistical Methods & Applications* 27(4):661–666
- Cerioli A, Farcomeni A, Riani M (2019) Wild adaptive trimming for robust estimation and cluster analysis. *Scandinavian Journal of Statistics* 46(1):235–256
- Chang WC (1983) On Using Principal Components Before Separating a Mixture of Two Multivariate Normal Distributions. *Applied Statistics* 32(3):267
- Chiang LH, Pell RJ (2004) Genetic algorithms combined with discriminant analysis for key variable identification. *Journal of Process Control* 14(2):143–155
- Dan G (1997) Algorithms on strings, trees and sequences: computer science and computational biology. Cambridge, UK: Press Syndacate of the Univestity of Cambridge
- Dash M, Liu H (1997) Feature selection for classification. *Intelligent Data Analysis* 1(1-4):131–156
- Dean N, Murphy TB, Downey G (2006) 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
- Dotto F, Farcomeni A, García-Escudero LA, Mayo-Iscar A (2018) A reweighting approach to robust clustering. *Statistics and Computing* 28(2):477–493
- Emerson JW, Green WA, Schloerke B, Crowley J, Cook D, Hofmann H, Wickham H (2013) The generalized pairs plot. *Journal of Computational and Graphical Statistics* 22(1):79–91
- Farcomeni A (2014) Robust constrained clustering in presence of entry-wise outliers. *Technometrics* 56(1):102–111
- Fernández Pierna JA, Dardenne P (2007) Chemometric contest at ‘Chimiométrie 2005’: A discrimination study. *Chemometrics and Intelligent Laboratory Systems* 86(2):219–223
- Fernández Pierna JA, Volery P, Besson R, Baeten V, Dardenne P (2005) Classification of Modified Starches by Fourier Transform Infrared Spectroscopy Using Support Vector Machines. *Journal of Agricultural and Food Chemistry* 53(17):6581–6585
- Fop M, Murphy TB (2018) Variable selection methods for model-based clustering. *Statistics Surveys* 12:18–65

- Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 97(458):611–631
- Gallegos MT, Ritter G (2005) A robust method for cluster analysis. *The Annals of Statistics* 33(1):347–380
- García-Escudero LA, Gordaliza A, Matrán C, Mayo-Iscar A (2008) A general trimming approach to robust cluster Analysis. *The Annals of Statistics* 36(3):1324–1345
- García-Escudero LA, Gordaliza A, Matrán C, Mayo-Iscar A (2011) Exploring the number of groups in robust model-based clustering. *Statistics and Computing* 21(4):585–599
- Gordaliza A (1991) Best approximations to random variables based on trimming procedures. *Journal of Approximation Theory* 64(2):162–180
- Guyon I, Aliferis C, Others (2007) Causal feature selection. In: *Computational methods of feature selection*, Chapman and Hall/CRC, pp 79–102
- Hamming RW (1950) Error detecting and error correcting codes. *The Bell System Technical Journal* 29(2):147–160
- Indahl U, Næs T (2004) A variable selection strategy for supervised classification with continuous spectroscopic data. *Journal of Chemometrics* 18(2):53–61
- John GH, Kohavi R, Pfleger K (1994) Irrelevant Features and the Subset Selection Problem. In: *Machine Learning Proceedings 1994*, Elsevier, pp 121–129
- Kass RE (1993) Bayes Factors in Practice. *The Statistician* 42(5):551
- Kass RE, Raftery AE (1995) Bayes Factors. *Journal of the American Statistical Association* 90(430):773
- Kohavi R, John GH (1997) Wrappers for feature subset selection. *Artificial Intelligence* 97(1-2):273–324
- Krusińska E, Liebhart J (1988) Robust Selection of the Most Discriminative Variables in the Dichotomous Problem with Application to Some Respiratory Disease Data. *Biometrical Journal* 30(3):295–303
- Liu H, Motoda H (2007) *Computational methods of feature selection*. CRC Press
- Mardia KV, Kent JT, Bibby JM (1979) *Multivariate analysis*. Academic Press London; New York
- Maugis C, Celeux G, Martin-Magniette ML (2009a) Variable Selection for Clustering with Gaussian Mixture Models. *Biometrics* 65(3):701–709

- Maugis C, Celeux G, Martin-Magniette ML (2009b) Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics & Data Analysis* 53(11):3872–3882
- Maugis C, Celeux G, Martin-Magniette ML (2011) Variable selection in model-based discriminant analysis. *Journal of Multivariate Analysis* 102(10):1374–1387
- McLachlan GJ (1992) *Discriminant Analysis and Statistical Pattern Recognition*, Wiley Series in Probability and Statistics, vol 544. John Wiley & Sons, Inc., Hoboken, NJ, USA
- Murphy TB, Dean N, Raftery AE (2010) 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
- Neykov N, Filzmoser P, Dimova R, Neytchev P (2007) Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics and Data Analysis* 52(1):299–308
- Pacheco J, Casado S, Núñez L, Gómez O (2006) Analysis of new variable selection methods for discriminant analysis. *Computational Statistics and Data Analysis* 51(3):1463–1478
- Raftery A, Hoeting J, Volinsky C, Painter I, Yeung KY (2018) *BMA: Bayesian Model Averaging*
- Raftery AE, Dean N (2006) Variable selection for model-based clustering. *Journal of the American Statistical Association* 101(473):168–178
- Rand WM (1971) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66(336):846
- Reid LM, O’Donnell CP, Downey G (2006) Recent technological advances for the determination of food authenticity. *Trends in Food Science & Technology* 17(7):344–353
- Riani M, Atkinson AC, Cerioli A, Corbellini A (2019) Efficient robust methods via monitoring for clustering and multivariate data analysis. *Pattern Recognition* 88:246–260
- Ritter G (2014) *Robust Cluster Analysis and Variable Selection*. Chapman and Hall/CRC
- Rousseeuw PJ (1984) Least Median of Squares Regression. *Journal of the American Statistical Association* 79(388):871–880
- Rousseeuw PJ, Bossche WVD (2018) Detecting Deviating Data Cells. *Technometrics* 60(2):135–145

- Rousseeuw PJ, Driessen KV (1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41(3):212–223
- Saeyns Y, Inza I, Larranaga P (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics* 23(19):2507–2517
- Schwarz G (1978) Estimating the dimension of a model. *The Annals of Statistics* 6(2):461–464
- Scrucca L, Raftery AE (2018) clustvarsel : A Package Implementing Variable Selection for Gaussian Model-Based Clustering in R. *Journal of Statistical Software* 84(1)
- Scrucca L, Fop M, Murphy TB, Raftery AE (2016) mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal* 8(1):289–317
- Strehl A, Ghosh J (2002) Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research* 3(Dec):583–617
- Todorov V (2007) Robust selection of variables in linear discriminant analysis. *Statistical Methods and Applications* 15(3):395–407
- Wolters MA (2015) A Genetic Algorithm for Selection of Fixed-Size Subsets with Application to Design Problems. *Journal of Statistical Software* 68(Code Snippet 1)
- Yu L (2008) Feature selection for genomic data analysis. *Computational methods of feature selection* pp 337–353
- Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research* 5(Oct):1205–1224
- Zhu X, Wu X (2004) Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review* 22(3):177–210