Advances in Data Analysis and Classification (2022) 16:55–92
https://doi.org/10.1007/s11634-021-00474-3

REGULAR ARTICLE

# Unobserved classes and extra variables in high-dimensional discriminant analysis

**Michael Fop[1] · Pierre-Alexandre Mattei[2] · Charles Bouveyron[2] · Thomas Brendan Murphy[3]**

Received: 29 January 2021 / Revised: 15 July 2021 / Accepted: 3 October 2021 / Published online: 1 March 2022


**Abstract**
In supervised classification problems, the test set may contain data points belonging to classes not observed in the learning phase. Moreover, the same units in the test data may be measured on a set of additional variables recorded at a subsequent stage with respect to when the learning sample was collected. In this situation, the classifier built in the learning phase needs to adapt to handle potential unknown classes and the extra dimensions. We introduce a model-based discriminant approach, Dimension-Adaptive Mixture Discriminant Analysis (D-AMDA), which can detect unobserved classes and adapt to the increasing dimensionality. Model estimation is carried out via a full inductive approach based on an EM algorithm. The method is then embedded in a more general framework for adaptive variable selection and classification suitable for data of large dimensions. A simulation study and an artificial experiment related to classification of adulterated honey samples are used to validate the ability of the proposed framework to deal with complex situations.

**Keywords** Adaptive supervised classification · Conditional estimation · Model-based discriminant analysis · Unobserved classes · Variable selection

**Mathematics Subject Classification** 62H30

✉ Michael Fop
michael.fop@ucd.ie

1 School of Mathematics & Statistics, University College Dublin, Dublin, Ireland

2 Université Côte d'Azur, Inria, CNRS, Laboratoire J.A. Dieudonné, Maasai team, Nice, France

3 School of Mathematics & Statistics and Insight Research Centre, University College Dublin, Dublin, Ireland

## 1 Introduction

Standard supervised classification approaches assume that all existing classes in the data have been observed during the learning phase. However, in some cases there could be the possibility of having units in the test set belonging to classes not previously observed. In such situation, a standard classifier would fail to detect the novel classes and would assign the observations only to the classes it is aware of from the learning stage. Moreover, the observations to be classified may be recorded on a collection of additional variables other than the variables already observed in the learning data. Examples of this situation are: classification of spectrometry data where the test data may be measured at a finer resolution than the learning set, hence with a increased number of wavelengths; classification of time-dependent data where variables correspond to points in time and observations are recorded in a continuous manner, whereby a given set of observations could have been collected up to a certain data point, while another set of units could have been recorded up to a successive period of time; classification of data where some of the variables of the training set are corrupted and cannot be used to build the classifier, while they are available in the testing stage. In all these scenarios, the classifier would also need to adapt to the increasing dimensionality. The combination of unrepresented classes in the training data and additional features in the test set leads to a complex situation where the model built in the learning stage is faced with two sources of criticality when classifying the new data: unobserved classes and extra variables.

In a recent work, Bouveyron (2014) introduced an adaptive method for model-based classification when the test data contains unknown classes. Nonetheless, the method is not capable of handling the situation of additional variables. To deal with this problem, this work introduces a model-based adaptive classification method for detection of novel classes in a set of new data that is characterized by an expanded number of variables with respect to the learning set. The approach is developed in conjunction with an adaptive variable selection procedure used to select the variables of the test set most relevant for the classification of the observations into observed and novel classes. An EM algorithm based on an inductive approach is proposed for estimation of the model. Variable selection is performed with a greedy forward algorithm that exploits the inductive characteristics of the approach and make it suitable for high-dimensional data.

The methodology presented here aims at tackling the problems arising from a mismatch in the distributions of labels and input variables in training and test data. This problem is more generally denoted as "dataset shif", and we point the interested reader to Quionero-Candela et al. (2009) and Moreno-Torres et al. (2012). In this work, the mismatch is due to unrepresented classes in the training data and increased dimensions of the test data.

### 1.1 Model-based discriminant analysis

Consider a set of learning observations $\{\mathbf{x}_s; \bar{\boldsymbol{\ell}}_s\}$, where $\mathbf{x}_s$ is the observation of a vector of random variables and $\bar{\boldsymbol{\ell}}_s$ is the associate class label, such that $\bar{\ell}_{sc} = 1$ if observation

$s$ belongs to class $c$, 0 otherwise; $c = 1, \ldots, C$. The aim of supervised classification is to build a classifier from the complete learning data $\{\mathbf{x}_s, \bar{\boldsymbol{\ell}}_s\}$ and use it to assign a new observation to one of the known classes. *Model-based discriminant analysis* (MDA, Bouveyron et al. 2019; McLachlan 2012, 2004; Fraley and Raftery 2002) is a probabilistic approach for supervised classification of continuous data in which the data generating process is represented as follows:

$$\begin{aligned} \bar{\boldsymbol{\ell}}_s &\sim \prod_{c=1}^{C} \tau_c^{\bar{\ell}_{sc}}, \\ (\mathbf{x}_s \mid \bar{\ell}_{sc} = 1) &\sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), \end{aligned} \tag{1}$$

where $\tau_c$ denotes the probability of observing class $c$, with $\sum_c \tau_c = 1$. Consequently, the marginal density of each data point corresponds to the density of a Gaussian mixture distribution:

$$f(\mathbf{x}_s\,;\,\boldsymbol{\Theta}) = \sum_{c=1}^{C} \tau_c\, \phi(\mathbf{x}_s\,;\,\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c),$$

where $\phi(.\,;\,\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ is the multivariate Gaussian density, with mean $\boldsymbol{\mu}_c$ and covariance matrix $\boldsymbol{\Sigma}_c$, and $\boldsymbol{\Theta}$ is the collection of all mixture parameters. Then, using the *maximum a posteriori* (MAP) rule, a new observation $\mathbf{y}_i$ is assigned to the class $\ell_{ic}$ with the highest posterior probability:

$$\Pr(\ell_{ic} = 1 \mid \mathbf{y}_i) = \frac{\tau_c\, \phi(\mathbf{y}_i\,;\,\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)}{\sum_{c=1}^{C} \tau_c\, \phi(\mathbf{y}_i\,;\,\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)}. \tag{2}$$

The framework is closely related to other discriminant analysis methods. If the covariance matrices are constrained to be the same across the classes, then the standard linear discriminant analysis (LDA) is recovered. On the other hand, if the covariance matrices have no constraints, the method corresponds to the standard quadratic discriminant analysis (QDA McLachlan 2004; Fraley and Raftery 2002). Several extension of this framework have been proposed in the literature in order to increase its flexibility and scope. For example, Hastie and Tibshirani (1996) consider the case where each class density is itself a mixture of Gaussian distributions with common covariance matrix and known number of components. Fraley and Raftery (2002) further generalize this approach, allowing the covariance matrices to be different across the sub-groups and applying model-based clustering to the observations of each class. Another approach, eigenvalue decomposition discriminant analysis (EDDA, Bensmail and Celeux 1996), is based on the family of parsimonious Gaussian models of Celeux and Govaert (1995), which imposes cross-constraints on the eigen-decomposition of the class covariance matrices. This latter approach allows more flexibility than LDA, and is more structured than QDA and the methods of Fraley and Raftery (2002), which could be over-parameterized. In high-dimensional settings, different approaches have been proposed based on regularization and variable selection: Friedman (1989) and Xu

et al. (2009) propose regularized versions of discriminant analysis where a shrinkage parameter is introduced to control the degree of regularization between LDA and QDA; Le et al. (2020) and Sun and Zhao (2015) define frameworks where a penalty term is introduced and the classes are characterized by sparse inverse covariance matrices. It is also worth to mention that for high-dimensional data, the framework of discriminant analysis has often been phrased in terms of sparse discriminant vectors, see for example: Clemmensen et al. (2011), Mai et al. (2012), Safo and Ahn (2016), Jiang et al. (2018), Qin (2018).

### 1.2 Adaptive mixture discriminant analysis

The discriminant analysis approaches pointed out earlier assume that all existing classes have been observed in the training set during the learning phase, not taking into account that the test data might include observations arising from classes present in the learning phase. Initial works in the the context of unobserved classes detection and model-based discriminant analysis are those of Miller and Browning (2003) and Frame and Jammalamadaka (2007), while examples of applications include galaxy classification (Bazell and Miller 2005) and acoustic species classification (Woillez et al. 2012). More recently, building on Miller and Browning (2003) work, Bouveyron (2014) introduced *Adaptive Mixture Discriminant Analysis* (AMDA), a framework for model-based discriminant analysis which allows the modeling of data where the test set contains novel classes not observed in the learning phase. The AMDA model considers the data arising from a mixture model with observed and unobserved classes and Bouveyron (2014) proposes two alternative approaches for model estimation. In particular, the *inductive* approach, where the classifying function is first estimated on the learning set and then applied to the test data. Crucially, the core assumption of the inductive approach is that the parameters estimated on the training data are fixed when dealing with the test set (see Chapelle et al. 2006; Pang and Kasabov 2004, for example). The assumption makes the approach most suitable for fast on-line data classification when the data come in multiple streams. In fact, with this approach, the learning set does not need to be kept in memory for prediction on a set of new data points, only the estimated parameters need to be stored.

In what follows we provide a formal description of the problem of unobserved classes in the test data and give a brief overview of the inductive AMDA methodology, as it constitutes the starting block of the main contribution of this paper. The learning data is composed of $M$ observations $\mathbf{x}_s$ and the associated class labels $\bar{\boldsymbol{\ell}}_s$, while the test data contains $N$ new observations $\mathbf{y}_i$. For ease of presentation, we treat the classes as sets, with $\mathcal{C}$ the set of all classes. The AMDA framework considers the situation where the data generating process is the same as depicted in (1) but only a subset of classes is observed in the training set, that is a subset $\mathcal{K} \subseteq \mathcal{C}$ of classes has been represented in the learning data. Therefore the test data may contain a set of extra "hidden" classes $\mathcal{H}$ such that $\mathcal{K} \cup \mathcal{H} = \mathcal{C}$. The cardinality of these sets (i.e. the number of classes) is denoted with $K$, $H$, and $C$ respectively, such as $K + H = C$.

The inductive AMDA approach consists of two phases: a learning phase and a discovery phase. The initial *learning phase* corresponds to the estimation of a model-

based discriminant analysis classifier using the training data. The data in the learning phase are complete, and the parameters estimated in this stage are then employed in the subsequent discovery phase. The *discovery phase* searches for $H$ novel classes in the set of new observations $\mathbf{y}_i$. In this phase, because of the inductive approach, the learning data is no longer needed and is discarded. The only relevant quantities to be retained are the parameter estimates obtained during the learning phase. In this stage, one needs to estimate the parameters of the unobserved classes in a partially unsupervised way in order to derive the classification rule as in (2). Since the observations $\mathbf{y}_i$ are unlabelled, the following log-likelihood is considered:

$$L(\mathbf{Y}; \boldsymbol{\Theta}) = \sum_{i=1}^{N} \log \left\{ \sum_{k=1}^{K} \tau_k \, \phi(\mathbf{y}_i \,;\, \bar{\boldsymbol{\mu}}_k, \bar{\boldsymbol{\Sigma}}_k) + \sum_{h=K+1}^{C} \tau_h \, \phi\left(\mathbf{y}_i \,;\, \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h\right) \right\},$$

where $\boldsymbol{\Theta}$ denotes the collection of all parameters. The parameters $\bar{\boldsymbol{\mu}}_k, \bar{\boldsymbol{\Sigma}}_k$ for $k = 1, \ldots, K$ are those of the classes observed in the training set and have been already estimated in the learning phase; the bar in the notation indicates that at this stage these parameters have already been estimated and are fixed. On the other hand, the Gaussian density parameters $\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h$ for $h = K + 1, \ldots, C$ remain to be estimated. Note that quantities related to the known classes are denoted with subscript $k$, while the subscript $h$ denotes quantities related to the new classes; subscript $c$ denotes both known and unknown classes. Bouveyron (2014) presents an EM algorithm (Dempster et al. 1977; McLachlan and Krishnan 2008) for optimization of the above log-likelihood with respect to the parameters of the unobserved classes, keeping fixed the parameters estimated in the learning phase.

### 1.3 Contribution and organization of the paper

The present paper extends the inductive AMDA framework to the case where the test data includes not only unobserved classes, but also extra variables. The contribution of this work is twofold. First, we propose a novel inductive model-based adaptive classification framework which can model the situation where the observations of the test data may contain classes unobserved during the training stage and are recorded on an expanded set of input features. Secondly, we incorporate this framework in a computationally efficient inductive variable selection procedure employed to detect the most relevant variables for classification into observed and unknown classes.

The paper is organized as follows. The current Sect. 1 introduced the problem of classifcation with unknown classes and extra variables, also providing a short overview of model-based classification via discriminant analysis. In particular, Sect. 1.2 briefly described the adaptive discriminant analysis method which is the basis of our proposed method. The following sections presents the novel methodology. Section 2 introduces the novel adaptive mixture discriminant analysis method capable to handle the complex situation where the new observations include information about unknown classes and are also measured on a set of additional variables. Section 3 presents an efficient inductive model estimation approach, based on a novel *inductive conditional estima-*

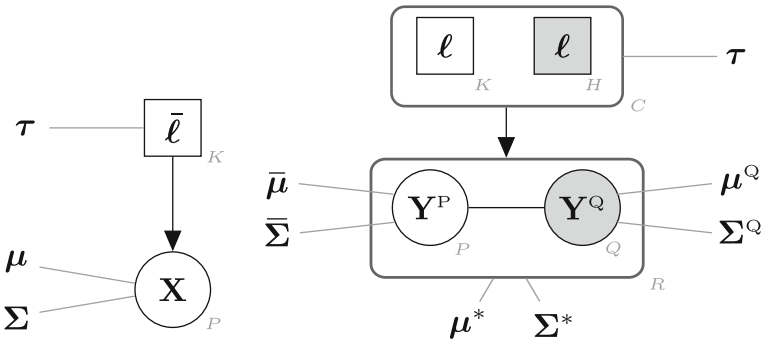
*tion* procedure employed to infer the parameters of unknown classes and unobserved variables. Technical details about initialization of the algorithm, assessing convergence, and model selection are described at the end of this section. In Sect. 4, the proposed method is naturally incorporated in a variable selection approach tailored for classification of high-dimensional data. Extensive simulation experiments are conducted in Sect. 5, in order to evaluate the performance of the proposed method for adaptive classification and variable selection. Section 6 presents an application to the classification of spectroscopy data of contaminated honey samples. The paper ends with a discussion in Sect. 7.

## 2 Dimension-adaptive mixture discriminant analysis

The AMDA framework combines supervised and unsupervised learning for detecting unobserved classes in the test data. However, in a dynamic classification setting, the new observations could be characterized not only by information about novel classes. In fact, the units in the test data could also have been recorded on a set of additional variables other than the ones already observed in the learning data. Typical examples are situations where the samples in the test data are collected at a finer resolution compared to the training set, some features are not available or corrupted during the training phase, or the training samples have been collected only up to a certain time point while the test set includes measurements concerning also subsequent time points.

Formally, we describe the setting of unknown classes and extra variables as follows. The learning data $\mathbf{X}$ is composed of $M$ observations $\mathbf{x}_s$ with the associated class labels $\bar{\boldsymbol{\ell}}_s$, and the test data $\mathbf{Y}$ is composed of $N$ new unlabelled observations $\mathbf{y}_i$. As in Sect. 1.2, the test data may contain a set of unobserved classes $\mathcal{H}$ such that $\mathcal{K} \cup \mathcal{H} = \mathcal{C}$. However, in this setting we consider the case where only a subset of variables available in the test data are observed or recorded in the training data. Hence, the test data also includes extra variables compared to the data set used for training. We consider the collection of variables observed in learning and test data as sets. In the case where only a subset of variables is available in the learning set, the test observations $\mathbf{y}_i$ are realizations of the set of variables $\mathcal{R}$, while the training observations are recorded on the subset of variables $\mathcal{P} \subset \mathcal{R}$. Consequently, the set $\mathcal{Q} = \mathcal{R} \setminus \mathcal{P}$ denotes the set of additional variables observed in the test set but not in the training set. The cardinalities of these sets, i.e. the number of variables in each set, are indicated with $P$, $Q$, and $R$, respectively, with $R = P + Q$.

The extra dimensions in the test data induce an augmented parameter space in the prediction and novel class detection stage of the classifier. Discarding the additional dimensions available in the test data can potentially damage the classification performance of the model, especially if the extra variables contain useful discriminant information. In this context, the classifier built in the learning phase needs to adapt in order to handle the situation where the new data to be classified contains information about novel classes and extra variables. To the purpose, we introduce *Dimension-Adaptive Mixture Discriminant Analysis* (D-AMDA). The model is a generalization of AMDA and is designed to classify new observations measured on additional variables and possibly containing information about unobserved classes. Under the model,




**(a)** *Learning phase.* **(b)** *Discovery phase.*

**Fig. 1** General framework of the inductive estimation approach for Dimension-Adaptive Mixture Discriminant Analysis

the joint densities of each observed and new data point together with observed and unobserved class labels are given by:

$$f(\mathbf{x}_s, \bar{\boldsymbol{\ell}}_s\,;\boldsymbol{\Theta}_x) = \prod_{k=1}^{K} \left\{\tau_k\,\phi(\mathbf{x}_s\,;\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right\}^{\bar{\ell}_{sk}}, \tag{3}$$

$$f(\mathbf{y}_i, \boldsymbol{\ell}_i\,;\boldsymbol{\Theta}_y) = \left[\prod_{k=1}^{K} \left\{\tau_k\,\phi(\mathbf{y}_i\,;\boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k^*)\right\}^{\ell_{ik}}\right] \times \left[\prod_{h=K+1}^{C} \left\{\tau_h\,\phi(\mathbf{y}_i\,;\boldsymbol{\mu}_h^*, \boldsymbol{\Sigma}_h^*)\right\}^{\ell_{ih}}\right], \tag{4}$$

with $s = 1, \ldots, M$, $i = 1, \ldots, N$, and $\boldsymbol{\Theta}_x$ and $\boldsymbol{\Theta}_y$ are the set of parameters for training and test observations. As earlier, the subscript $k$ indicates quantities related to the known classes, while the subscript $h$ denotes quantities related to the new classes. The parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the class-specific mean and covariance parameters of the observed classes in the learning data and related to the subset of variables $\mathcal{P}$. The parameters denoted with $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$ denotes respectively the class-specific mean and covariance parameters for both observed and unobserved classes and related to the full collection of variables of the test data. These parameters are defined on an augmented space compared to the parameters in (3). Indeed, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are $P$-dimensional vectors and $P \times P$ matrices, while $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$ are $R$-dimensional vectors and $R \times R$ matrices. As such, the model takes into account that $\mathbf{y}_i$ may be measured on additional variables and generalizes the AMDA framework.

Similarly to AMDA, model estimation for D-AMDA is carried out within an inductive estimation framework. Figure 1 provides a sketch of the general framework. In (a) the training data $\mathbf{X}$ and the corresponding collection of labels $\bar{\boldsymbol{\ell}}$ are observed. The aim of the learning stage is to estimate the set of parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\tau}$ of the density in (3). In (b) only $\mathbf{Y}$ is observed and no information about the classification is given. The test data is partitioned into two parts: $\mathbf{Y}^{\mathrm{P}}$, the subset of data corresponding to the variables observed in the training set, and $\mathbf{Y}^{\mathrm{Q}}$, the subset of data related to the

additional variables (gray background). In the discovery phase the aim is to estimate the parameters $\boldsymbol{\mu}_k^*$, $\boldsymbol{\mu}_h^*$, $\boldsymbol{\Sigma}_k^*$, $\boldsymbol{\Sigma}_h^*$ and $\boldsymbol{\tau}$ of the distribution in (4), as well as to infer the classification of the new unlabelled data points. The collection of class labels to be inferred here is composed of the labels indicating the classes observed in the learning stage and the labels indicating the new classes (gray background). Model estimation for D-AMDA is detailed in the next sections.

## 3 Inductive model estimation and inference

The use of an inductive estimation approach is appropriate for the proposed D-AMDA framework, as it allows to only retain the test data once a set of mean and covariance parameter estimates have been obtained from the training data. Since the training set is lower dimensional compared to the test data, this can be particularly efficient in high-dimensional and on-line classification settings. Like in Sect. 1.2, the approach is composed of a learning and a discovery phase. In this case, the discovery phase includes a novel estimation procedure employed to account for the extra dimensions.

### 3.1 Learning phase

The learning phase of the inductive approach consists of estimating parameters for the observed classes employing only the training set. From Eq. (3), this stage corresponds to the standard estimation of a model-based discriminant analysis classifier, performed by optimization of the associated log-likelihood:

$$L(\mathbf{X}, \bar{\boldsymbol{\ell}}\,; \boldsymbol{\Theta}_x) = \sum_{s=1}^{M} \sum_{k=1}^{K} \bar{\ell}_{sk} \log \left\{ \tau_k \, \phi(\mathbf{x}_s\,; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\},$$

which reduces to the separate estimation of the class density parameters. Here we consider the eigenvalue decomposition discriminant analysis (EDDA) of Bensmail and Celeux (1996), in which the class covariance matrices $\boldsymbol{\Sigma}_k$ are parameterized according to the eigen-decomposition $\boldsymbol{\Sigma}_k = \lambda_k \, \mathbf{D}_k \, \mathbf{A}_k \, \mathbf{D}_k^{'}$, providing a collection of parsimonious models. Estimation of this model is carried out using the `mclust` R package Scrucca et al. (2016), which also automatically selects the best covariance decomposition model using the *Bayesian Information Criterion* (BIC, Schwarz 1978; Fraley and Raftery 2002). This learning phase is more general than the one in Bouveyron (2014). In fact, the author considers a QDA model in the learning phase, a particular case of EDDA corresponding to an unconstrained covariance model (see Scrucca et al. 2016). The EDDA classifier learned in this phase is more flexible and is proven to perform better than QDA (Bensmail and Celeux 1996), although it will introduce some complications, which described in the following section.

The learning phase outputs the parameters of the EDDA model fitted on the training data $\bar{\boldsymbol{\mu}}_k$ and $\bar{\boldsymbol{\Sigma}}_k$ for $k = 1, \ldots, K$. We note again that we use the bar symbol to stress the fact that the parameters estimated in the learning phase are fixed during the discovery

phase. Since the discovery phase relies only on the test data and these parameters, the training set can be discarded.

### 3.2 Discovery phase

The discovery phase looks for novel classes in the test data, given the parameter estimates from the learning phase. Under the D-AMDA modelling framework, we also need to take into account the extra dimensions of the test data. Subsequently, in this phase we need to estimate two main collections of parameters: the parameters of the additional variables corresponding to novel and known classes, and the parameters of the already observed variables related to new and known classes. These characterize the distribution in (4) and are estimated keeping the parameter estimates from the learning phase fixed.

Because the labels of the test data are unobserved, in this stage we aim to optimize the following log-likelihood:

$$L(\mathbf{Y};\boldsymbol{\Theta}_y)=\sum_{i=1}^{N}\log\left\{\sum_{k=1}^{K}\tau_k\,\phi(\mathbf{y}_i\,;\boldsymbol{\mu}_k^*,\boldsymbol{\Sigma}_k^*)+\sum_{h=K+1}^{C}\tau_h\,\phi(\mathbf{y}_i\,;\boldsymbol{\mu}_h^*,\boldsymbol{\Sigma}_h^*)\right\}, \tag{5}$$

with $\sum_{k=1}^{K}\tau_k+\sum_{h=1}^{H}\tau_h=1$. Crucially, for $k=1,\ldots,K$, mean and covariance parameters of the $K$ observed classes are partitioned into parameters fixed from the learning phase corresponding to the variables observed in $\mathbf{X}$ and parameters corresponding to the additional variables present in the test data:

$$\boldsymbol{\mu}_k^*=(\bar{\boldsymbol{\mu}}_k\;\;\boldsymbol{\mu}_k^{\mathrm{Q}})'\qquad\boldsymbol{\Sigma}_k^*=\begin{bmatrix}\bar{\boldsymbol{\Sigma}}_k & \mathbf{C}_k\\ \mathbf{C}_k' & \boldsymbol{\Sigma}_k^{\mathrm{Q}}\end{bmatrix}, \tag{6}$$

where $\mathbf{C}_k$ are the covariance terms between additional and observed variables. Such partition of the parameters of the observed classes will need to be taken into account during the estimation procedure, as it indirectly induces a constraint on the estimation of the parameters for the additional variables; see the following Sect. 3.2.2.

Optimization of this log-likelihood in the discovery phase is carried out by resorting to an EM algorithm (Dempster et al. 1977; McLachlan and Krishnan 2008). From (5) we have the complete log-likelihood:

$$\begin{aligned}L(\mathbf{Y},\boldsymbol{\ell};\boldsymbol{\Theta}_y)=\sum_{i=1}^{N}\Bigg[&\sum_{k=1}^{K}\ell_{ik}\log\Big\{\tau_k\,\phi(\mathbf{y}_i\,;\boldsymbol{\mu}_k^*,\boldsymbol{\Sigma}_k^*)\Big\}\\&+\sum_{h=K+1}^{C}\ell_{ih}\log\Big\{\tau_h\,\phi(\mathbf{y}_i\,;\boldsymbol{\mu}_h^*,\boldsymbol{\Sigma}_h^*)\Big\}\Bigg],\end{aligned} \tag{7}$$

where, $\ell_{ik}$ and $\ell_{ih}$ denote the latent class membership indicators to be estimated on the test data for known and unobserved classes. The EM algorithm alternates the following two steps.

- **E Step**: After estimation of the parameters at the previous M step iteration, the estimated conditional probabilities, $t_{ic} = \widehat{\Pr}(\ell_{ic} = 1 \mid \mathbf{y}_i)$ are computed as:

$$t_{ic} = \frac{\widehat{\tau}_c \, \phi(\mathbf{y}_i ; \widehat{\boldsymbol{\mu}}_k^*, \widehat{\boldsymbol{\Sigma}}_k^*)}{\sum_{k=1}^{K} \widehat{\tau}_k \, \phi(\mathbf{y}_i ; \widehat{\boldsymbol{\mu}}_k^*, \widehat{\boldsymbol{\Sigma}}_k^*) + \sum_{h=K+1}^{C} \widehat{\tau}_h \, \phi(\mathbf{y}_i ; \widehat{\boldsymbol{\mu}}_h^*, \widehat{\boldsymbol{\Sigma}}_h^*)},$$

  for $i = 1, \ldots, N$, and $c = 1, \ldots, C$.
- **M Step**: in this step of the algorithm we maximize the expectation of the complete log-likelihood computed using the estimated probabilities $t_{ic}$ of the E step. Due to the augmented dimensions of the test data, this step is more involved. The optimization procedure of the M-step is divided into two parts: estimation of mixing proportions and mean and covariance parameters corresponding to the unobserved classes, described in Sect. 3.2.1, and estimation of mean and covariance parameters related to the classes already observed in the learning phase, described in Sect. 3.2.2.

### 3.2.1 Estimation of mixing proportions and parameters of unobserved classes

The introduction of new variables does not affect the estimation of the mixing proportions, nor the estimation of the parameters corresponding to the new classes. Hence, in this case the updates are in line with those outlined in Bouveyron (2014). From Eq. (7), the estimates of mean and covariance parameters of the $H$ hidden classes are obtained simply by optimizing the term involving $\boldsymbol{\mu}_h^*$ and $\boldsymbol{\Sigma}_h^*$. Therefore, the estimates of the Gaussian density parameters related to the unknown classes are simply given by:

$$\widehat{\boldsymbol{\mu}}_h^* = \frac{1}{N_h} \sum_{i=1}^{N} t_{ih} \mathbf{y}_i, \qquad \widehat{\boldsymbol{\Sigma}}_h^* = \frac{1}{N_h} \sum_{i=1}^{N} t_{ih} (\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_h^*)(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_h^*)',$$

with $N_h = \sum_i t_{ih}$. For the mixing proportions, two alternative updates are available. One is based on the re-normalization of the mixing proportions $\bar{\tau}_k$ as outlined in Bouveyron (2014), the other on the re-estimation of the mixing proportions for both observed and unobserved classes on the test data. The two updates correspond to very different assumptions about the data: the re-normalization update is based on the assumption that the new classes do not affect the balance of the classes observed in the training set, while the other update is based on the assumption that the class proportions may have changed in the test data. We opt for this latter approach, since it is more flexible and avoids the introduction of possible bias due to the re-normalization, as discussed in Lawoko and McLachlan (1989). The mixing proportions are updated as follows:

$$\widehat{\tau}_k = \frac{N_k}{N} \qquad \widehat{\tau}_h = \frac{N_h}{N} \quad \text{for} \quad k = 1, \ldots, K \quad \text{and} \quad h = K+1, \ldots, C.$$

### 3.2.2 Inductive conditional estimation procedure

The estimation of mean and covariance parameters $\boldsymbol{\mu}_k^*$ and $\boldsymbol{\Sigma}_k^*$ of the classes already observed in the training data is an involved problem, due to the augmented parameter dimensions and the fact that the parameters from the learning phase need to be kept fixed. Here we need to estimate the components $\boldsymbol{\mu}_k^{\mathrm{Q}}$, $\boldsymbol{\Sigma}_k^{\mathrm{Q}}$ and $\mathbf{C}_k$ of the partitions in (6). As in a standard Gaussian mixture model, a straightforward update for these would be computing the related in sample weighted quantities. However, this would not take into account the constraint that the parameters $\bar{\boldsymbol{\mu}}_k$ and $\overline{\boldsymbol{\Sigma}}_k$ have already been estimated in the learning phase and need to be held fixed. In particular, the covariance block $\overline{\boldsymbol{\Sigma}}_k$ has been estimated in the learning phase via the EDDA model, imposing constraints on its eigen-decomposition. As it is often the case, if the covariance model for $\overline{\boldsymbol{\Sigma}}_k$ has a particular structure (i.e. is not the VVV using the `mclust` nomenclature) the approach would not ensure a valid positive definite $\boldsymbol{\Sigma}_k^*$. A clear example is the case where the EDDA model estimated in the learning phase is a spherical one with diagonal matrices $\overline{\boldsymbol{\Sigma}}_k$. In such case, completing the off-diagonal entries of $\boldsymbol{\Sigma}_k^*$ with non-zero terms and without taking into account the structure of the block $\overline{\boldsymbol{\Sigma}}_k$ would not guarantee a positive definite covariance matrix (Zhang 2006). We propose the following procedure to obtain valid estimates.

Denote an observation of the test data $\mathbf{y}_i = \{\mathbf{y}_i^{\mathrm{P}}, \mathbf{y}_i^{\mathrm{Q}}\}$, where $\mathbf{y}_i^{\mathrm{P}}$ are the measurements of the set $\mathcal{P}$ of variables of the training data and $\mathbf{y}_i^{\mathrm{Q}}$ are the measurements of the set $\mathcal{Q}$ of additional variables observed in the test data. To take into account the structure of the block $\overline{\boldsymbol{\Sigma}}_k$, the problem of maximizing the expectation of (7) with respect to $\boldsymbol{\Sigma}_k^*$ and $\boldsymbol{\mu}_k^*$ can be interpreted as the problem of finding estimates of $\boldsymbol{\mu}_k^{\mathrm{Q}}$, $\boldsymbol{\Sigma}_k^{\mathrm{Q}}$, and $\mathbf{C}_k$ such that the joint distribution of observed and extra variables $(\{\mathbf{y}_i^{\mathrm{P}}, \mathbf{y}_i^{\mathrm{Q}}\} \,|\, l_{ik} = 1) \sim \mathcal{N}(\boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k^*)$ is a multivariate Gaussian density whose marginal distributions are $(\mathbf{y}_i^{\mathrm{P}} \,|\, l_{ik} = 1) \sim \mathcal{N}(\bar{\boldsymbol{\mu}}_k, \overline{\boldsymbol{\Sigma}}_k)$ and $(\mathbf{y}_i^{\mathrm{Q}} \,|\, l_{ik} = 1) \sim \mathcal{N}(\boldsymbol{\mu}_k^{\mathrm{Q}}, \boldsymbol{\Sigma}_k^{\mathrm{Q}})$, and with $\boldsymbol{\Sigma}_k^*$ being positive definite. To accomplish this task, we devise the following *inductive conditional estimation* procedure:

*Step*1. Fix the marginal distribution of the variables observed in the learning phase, $(\mathbf{y}_i^{\mathrm{P}} \,|\, l_{ik} = 1) \sim \mathcal{N}(\bar{\boldsymbol{\mu}}_k, \overline{\boldsymbol{\Sigma}}_k)$;

*Step*2. Estimate the parameters of the conditional distribution $(\mathbf{y}_i^{\mathrm{Q}} \,|\, \mathbf{y}_i^{\mathrm{P}}, l_{ik} = 1) \sim \mathcal{N}(\mathbf{m}_{ik}, \mathbf{E}_k)$, where $\mathbf{m}_{ik}$ and $\mathbf{E}_k$ are related mean and covariance parameters.

*Step*3. Find estimates of the parameters of the joint distribution $(\{\mathbf{y}_i^{\mathrm{P}}, \mathbf{y}_i^{\mathrm{Q}}\} \,|\, l_{ik} = 1) \sim \mathcal{N}(\boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k^*)$ using the fixed marginal and the conditional distribution.

Since we are using an inductive approach, *Step 1* corresponds in keeping $\bar{\boldsymbol{\mu}}_k$, $\overline{\boldsymbol{\Sigma}}_k$ fixed. Next, in *Step 2* the parameter estimates of the distribution of the new variables given the variables observed in the training set are obtained. This allows to take into account the information and the structure of the learning phase parameters. Then, in *Step 3* these estimates are used to find the parameters of the marginal distribution of the set of new variables $\mathcal{Q}$ and the joint distribution of $\mathcal{R} = \{\mathcal{P}, \mathcal{Q}\}$, while preserving the joint association structure among all the variables in $\mathcal{R}$. The proposed method is related to the well known *iterative proportional fitting* algorithm for fitting distributions with fixed marginals (see for example Whittaker 1990; Fienberg and Meyer 2006), and the

*iterative conditional fitting* algorithm of Chaudhuri et al. (2007) used to estimate a multivariate Gaussian distribution with association constraints.

Taking the expectation of the complete log-likelihood in (7), the term involving $\boldsymbol{\mu}_k^*$ and $\boldsymbol{\Sigma}_k^*$ can be rewritten as:

$$\sum_{i=1}^{N}\left[\sum_{k=1}^{K} t_{ik}\log\left\{\phi(\mathbf{y}_i^{\mathrm{Q}}\mid\mathbf{y}_i^{\mathrm{P}};\,\mathbf{m}_{ik},\mathbf{E}_k)\,\phi(\mathbf{y}_i^{\mathrm{P}};\,\bar{\boldsymbol{\mu}}_k,\bar{\boldsymbol{\Sigma}}_k)\right\}\right]. \tag{8}$$

In *Step 1*, parameters $\bar{\boldsymbol{\mu}}_k, \bar{\boldsymbol{\Sigma}}_k$ are fixed from the learning phase. Therefore, the term $\log\{\phi(\mathbf{y}_i^{\mathrm{P}};\,\bar{\boldsymbol{\mu}}_k,\bar{\boldsymbol{\Sigma}}_k)\}$ is already maximized. In *Step 2* and *Step 3*, we make use of the well known closure properties of the multivariate Gaussian distribution (see Tong 1990; Zhang 2006, for example) in order to maximize the term $\log\{\phi(\mathbf{y}_i^{\mathrm{Q}}\mid\mathbf{y}_i^{\mathrm{P}};\,\mathbf{m}_k,\mathbf{E}_k)\}$. In *Step 2* the focus is on the conditional distribution; for each observation $i$ we can rewrite:

$$\mathbf{m}_{ik}=\boldsymbol{\mu}_k^{\mathrm{Q}}+\mathbf{C}_k'\bar{\boldsymbol{\Sigma}}_k^{-1}(\mathbf{y}_i^{\mathrm{P}}-\bar{\boldsymbol{\mu}}_k),\qquad \mathbf{E}_k=\boldsymbol{\Sigma}_k^{\mathrm{Q}}-\mathbf{C}_k'\,\bar{\boldsymbol{\Sigma}}_k^{-1}\,\mathbf{C}_k.$$

Let us define the scattering matrix $\mathbf{O}_k=\sum_{i=1}^{N} t_{ik}(\mathbf{y}_i-\bar{\mathbf{y}}_k)(\mathbf{y}_i-\bar{\mathbf{y}}_k)'$, with $\bar{\mathbf{y}}_k=\frac{1}{N_k}\sum_{i=1}^{N} t_{ik}\,\mathbf{y}_i$. We can partition it as:

$$\mathbf{O}_k=\begin{bmatrix}\mathbf{W}_k & \mathbf{V}_k\\ \mathbf{V}_k' & \mathbf{U}_k\end{bmatrix},$$

with $\mathbf{W}_k$ the block related to the variables observed in the learning set, $\mathbf{U}_k$ the block associated to the new variables and $\mathbf{V}_k$ the crossproducts. Now we maximize (8) with respect to $\mathbf{E}_k$ and $\mathbf{C}_k$. After some algebraic manipulations, we obtain the estimates:

$$\begin{aligned}\widehat{\mathbf{C}}_k&=(\bar{\boldsymbol{\Sigma}}_k^{-1}\,\mathbf{W}_k\,\bar{\boldsymbol{\Sigma}}_k^{-1})^{-1}(\bar{\boldsymbol{\Sigma}}_k^{-1}\mathbf{V}_k),\\ \widehat{\mathbf{E}}_k&=\frac{1}{N_k}\left[\widehat{\mathbf{C}}_k'\,\bar{\boldsymbol{\Sigma}}_k^{-1}\,\mathbf{W}_k\,\bar{\boldsymbol{\Sigma}}_k^{-1}\,\widehat{\mathbf{C}}_k-2\mathbf{V}_k'\bar{\boldsymbol{\Sigma}}_k^{-1}\,\widehat{\mathbf{C}}_k+\mathbf{U}_k\right].\end{aligned}$$

Then, in *Step 3* we obtain the estimates of the marginal distribution for the set of extra variables as:

$$\widehat{\boldsymbol{\mu}}_k^{\mathrm{Q}}=\frac{1}{N_k}\left[\sum_{i=1}^{N} t_{ik}\mathbf{y}_i^{\mathrm{Q}}-\widehat{\mathbf{C}}_k'\,\bar{\boldsymbol{\Sigma}}_k^{-1}\sum_{i=1}^{N} t_{ik}(\mathbf{y}_i^{\mathrm{P}}-\bar{\boldsymbol{\mu}}_k)\right],\qquad \widehat{\boldsymbol{\Sigma}}_k^{\mathrm{Q}}=\widehat{\mathbf{E}}_k+\widehat{\mathbf{C}}_k'\,\bar{\boldsymbol{\Sigma}}_k^{-1}\,\widehat{\mathbf{C}}_k.$$

Hence $\boldsymbol{\mu}_k^*$ and $\boldsymbol{\Sigma}_K^*$ are estimated:

$$\widehat{\boldsymbol{\mu}}_k^*=(\bar{\boldsymbol{\mu}}_k\;\;\widehat{\boldsymbol{\mu}}_k^{\mathrm{Q}})',\qquad \widehat{\boldsymbol{\Sigma}}_k^*=\begin{bmatrix}\bar{\boldsymbol{\Sigma}}_k & \widehat{\mathbf{C}}_k\\ \widehat{\mathbf{C}}_k' & \widehat{\boldsymbol{\Sigma}}_k^{\mathrm{Q}}\end{bmatrix}.$$

Further details about the derivations are in Appendix 1. Provided that $\mathbf{O}_k$ is positive definite, the estimate of $\boldsymbol{\Sigma}_k$ obtained in such way is ensured to be positive definite

as well due to the properties of the Schur complement (Zhang 2006; Tong 1990). In certain cases, for example when the number of variables $R$ is large compared to $N$ and to the expected class sizes, or when the variables are highly correlated, this scattering matrix could be singular. To overcome this issue, one could resort to regularization. To this purpose, we delineate a simple Bayesian regularization approach in Appendix 1.

### 3.3 Technical details

#### 3.3.1 Initialization of the EM algorithm

In order to compute the first E step iteration of the EM algorithm in the discovery phase, we need to initialize the parameter values. A random initialization has a fair chance of not providing good starting points. On the other hand, the initialization based on the model-based hierarchical clustering method discussed in (Scrucca and Raftery 2015) and (Fraley 1998) often yields good starting points, is computationally efficient and works well in practice. However, we need to take care of the fact that a subset of the parameters is fixed. We make use of the following strategy for initialization.

First we obtain a hierarchical unsupervised partition of the observations in the test data using the method of Scrucca and Raftery (2015) and Fraley (1998). Afterwards, for a fixed number $C$ of clusters and the corresponding partition, we compute the within-cluster means and covariance matrices, both for new and observed variables. Let us denote with $\tilde{\boldsymbol{\mu}}_g^{\mathrm{P}}$ and $\tilde{\boldsymbol{\Sigma}}_g^{\mathrm{P}}$ $(g = 1, \ldots, C)$ the computed cluster parameters related to the observed variables, with $\tilde{\boldsymbol{\mu}}_g^{\mathrm{Q}}$ and $\tilde{\boldsymbol{\Sigma}}_g^{\mathrm{Q}}$ those related to the extra variables, and with $\tilde{\mathbf{C}}_g$ the covariance terms. Now, we find which of the detected clusters match the classes observed in the training set over the observed variables. For each known class and each cluster we compute the Kullback-Leibler divergence:

$$\operatorname{tr}\left\{\left(\tilde{\boldsymbol{\Sigma}}_g^{\mathrm{P}}\right)^{-1}\overline{\boldsymbol{\Sigma}}_k\right\} + \left(\tilde{\boldsymbol{\mu}}_g^{\mathrm{P}} - \overline{\boldsymbol{\mu}}_k\right)'\left(\tilde{\boldsymbol{\Sigma}}_g^{\mathrm{P}}\right)^{-1}\left(\tilde{\boldsymbol{\mu}}_g^{\mathrm{P}} - \overline{\boldsymbol{\mu}}_k\right) + \log\frac{\det\tilde{\boldsymbol{\Sigma}}_g^{\mathrm{P}}}{\det\overline{\boldsymbol{\Sigma}}_k}, \qquad \forall\, g, k.$$

Then, we find the first $K$ clusters with the minimum divergence and thus likely corresponding to the classes observed in the training data. For these clusters, the set of parameters related to the observed variables are initialized with the associated values $\overline{\boldsymbol{\mu}}_k$ and $\overline{\boldsymbol{\Sigma}}_k$, the set of parameters related to the new variables are initialized with the same values $\tilde{\boldsymbol{\mu}}_k^{\mathrm{Q}}$ and $\tilde{\boldsymbol{\Sigma}}_k^{\mathrm{Q}}$, and the covariance terms with $\tilde{\mathbf{C}}_k$. The remaining clusters can be considered as hidden classes and the related parameters are initialized with the corresponding cluster means and covariances.

#### 3.3.2 Selection of the number of hidden classes

Similarly to AMDA, also in the D-AMDA framework class detection corresponds to selection of the number of hidden classes in the test data. As in the learning phase, the BIC is employed for this purpose. Explicitly, for a range of values of number of

hidden classes $H$, we choose the model that maximizes the quantity:

$$\mathrm{BIC}_H = 2\,L(\mathbf{Y}; \widehat{\boldsymbol{\Theta}}_y) - \eta_H \log N,$$

where $\eta_H$ is the number of parameters estimated in the discovery phase, equal to $(H + K - 1) + 2HR + H\binom{R}{2} + 2KQ + KPQ + K\binom{Q}{2}$.

### 3.3.3 Assessing convergence

To determine the convergence of the EM algorithm in the discovery phase we employ a standard stopping criterion, monitoring the relative change of the value of the maximized log-likelihood in (5). Let $L^{(t)}$ be the value of the objective function in (5) at iteration $t$. The EM algorithm is stopped when $\frac{|L^{(t)} - L^{(t-1)}|}{1+|L^{(t)}|} < \varepsilon$, where the tolerance $\varepsilon = 10^{-5}$.

## 4 Inductive variable selection for D-AMDA

Given the large amount and the variety of sources at disposition, classification of high-dimensional data is becoming more and more a routine task. In this setting, variable selection has been proven beneficial for increasing accuracy, reducing the number of parameters and a better model interpretation (Guyon and Elisseeff 2003; Pacheco et al. 2006; Brusco and Steinley 2011; Fop and Murphy 2018). We adapt the variable selection method of Maugis et al. (2011) and Murphy et al. (2010) in order to perform *inductive* variable selection within the context of D-AMDA. The aim is to select the relevant variables that contain the most useful information about both observed and novel classes. The method is inductive in the sense that the classifier model first is built on the data observed in the learning phase. Then, while performing variable selection on the new test data, the classifier is adapted by removing and adding variables without re-estimating the model on the learning data.

Following Maugis et al. (2011) and Murphy et al. (2010), at each step of the variable selection procedure we consider the partition $\mathbf{Y} = (\mathbf{Y}^{\text{class}}, Y^{\text{prop}}, \mathbf{Y}^{\text{other}})$, where $\mathbf{Y}^{\text{class}}$ is the current set of relevant variables, $Y^{\text{prop}}$ is the variable proposed to be added/removed to/from $\mathbf{Y}^{\text{class}}$, and $\mathbf{Y}^{\text{other}}$ are the non relevant variables. Let also $\boldsymbol{\ell}$ be the class indicator variable. For each stage of the algorithm, we compare two models:

$$\begin{aligned}
\mathcal{M}_1 &: \; p(\mathbf{Y} \mid \boldsymbol{\ell}) = p(\mathbf{Y}^{\text{class}}, Y^{\text{prop}} \mid \boldsymbol{\ell})\, p(\mathbf{Y}^{\text{other}}), \\
\mathcal{M}_2 &: \; p(\mathbf{Y} \mid \boldsymbol{\ell}) = p(\mathbf{Y}^{\text{class}} \mid \boldsymbol{\ell})\, p(Y^{\text{prop}} \mid \mathbf{Y}^{\text{reg}} \subseteq \mathbf{Y}^{\text{class}})\, p(\mathbf{Y}^{\text{other}}).
\end{aligned}$$

In model $\mathcal{M}_1$, $\mathbf{Y}^{\text{prop}}$ is relevant for classification and $p(\mathbf{Y}^{\text{class}}, Y^{\text{prop}} \mid \boldsymbol{\ell})$ is the D-AMDA model where the classifier is adapted by including the proposed variable $Y^{\text{prop}}$. In model $\mathcal{M}_2$, $Y^{\text{prop}}$ does not depend on the labels and thus is not useful for classification. $p(\mathbf{Y}^{\text{class}} \mid \boldsymbol{\ell})$ is the D-AMDA model on the current selected variables and the conditional distribution $p(Y^{\text{prop}} \mid \mathbf{Y}^{\text{reg}} \subseteq \mathbf{Y}^{\text{class}})$ is a regression where $Y^{\text{prop}}$

depends on $\mathbf{Y}^{\text{class}}$ through a subset of predictors $\mathbf{Y}^{\text{reg}}$. This regression term encompasses the fact that some variables may be redundant given the set of already selected ones, and thus can be discarded (Murphy et al. 2010; Raftery and Dean 2006). Relevant predictors are chosen via a standard stepwise procedure and the selection avoids to include unnecessary parameters that would over-penalize the model without a significant increase in its likelihood (Maugis et al. 2009a, b). The two models are compared by computing the difference between their BIC:

$$\begin{aligned}\text{BIC}_1 &= \text{BIC}_{\text{class}}(\mathbf{Y}^{\text{class}}, Y^{\text{prop}}),\\ \text{BIC}_2 &= \text{BIC}_{\text{no class}}(\mathbf{Y}^{\text{class}}) + \text{BIC}_{\text{reg}}(Y^{\text{prop}} \mid \mathbf{Y}^{\text{reg}} \subseteq \mathbf{Y}^{\text{class}}),\end{aligned}$$

where $\text{BIC}_{\text{class}}(\mathbf{Y}^{\text{class}}, Y^{\text{prop}})$ is the BIC of the D-AMDA model where $Y^{\text{prop}}$ is useful for classification, $\text{BIC}_{\text{no class}}(\mathbf{Y}^{\text{class}})$ is the BIC on the current set of selected variables and $\text{BIC}_{\text{reg}}(Y^{\text{prop}} \mid \mathbf{Y}^{\text{reg}} \subseteq \mathbf{Y}^{\text{class}})$ is the BIC of the regression. The difference $(\text{BIC}_1 - \text{BIC}_2)$ is computed and if it is greater than zero, there is evidence that $Y^{\text{prop}}$ conveys useful information about the classes, hence variable $Y^{\text{prop}}$ is added to the D-AMDA model and the classifier is updated.

The selection is performed using a stepwise greedy forward search where variables are added and removed in turn. Since we adopt an inductive approach, when the variables to be added/removed belong to the set of variables already observed in the learning phase, the classifier is updated in a fast and efficient way. Indeed, if a variable observed in $\mathbf{X}$ needs to be added, the classifier is updated by simply augmenting the set of parameters with the parameters already estimated in the learning phase. Analogously, if the variable needs to be removed, the classifier is updated by deleting the corresponding parameters. Only parameters related to additional variables and novel classes need to be estimated when updating the D-AMDA model. Parameters related to known classes and observed variables are updated only via deletion or addition. Moreover, the forward greedy search employed to add and remove variables can be easily separated into a collection of parallel model comparison tasks. Therefore, the variable selection procedure can be implemented exploiting parallel computing, which considerably reduces the computing time (see Scrucca and Raftery (2018) for a discussion on the advantages of parallel computing for variable selection for model-based clustering). As such, the method is suitable for fast on-line variable selection.

The classification procedure is partly unsupervised because of the presence of unobserved classes. Therefore, while searching for the relevant variables, also the number $H$ of unknown classes needs to be chosen. As in Maugis et al. (2009a, b); Raftery and Dean (2006), we consider a range of possible values for $H$. Then, at every step $\text{BIC}_{\text{class}}(\mathbf{Y}^{\text{class}}, Y^{\text{prop}})$ and $\text{BIC}_{\text{no class}}(\mathbf{Y}^{\text{class}})$ are computed by maximizing over this range. Therefore, the method returns both the set of relevant variables and the optimal number of unobserved classes.

The set of relevant variables needs to be initialized at the first stage of the variable selection algorithm. We suggest to start the search from a conveniently chosen subset of size $S$ of the variables observed in the learning phase. To determine such subset, for every variable in $\mathbf{Y}$ corresponding to those already observed in $\mathbf{X}$, we estimate a univariate Gaussian mixture model for a number of components ranging from 2 to

$G > K$. Then we compute the difference between the BIC of such model and the BIC of a single univariate Gaussian distribution. The variables are ranked according to this difference from the largest to the lowest value. The starting subset is formed by selecting the top $S$ variables in the list. Similar initial selection strategies have been discussed in McLachlan (2004) and Murphy et al. (2010). Note that one could also initialize the set of relevant variables from all the observed variables of the training data. Nonetheless, if the number of variables observed in $\mathbf{X}$ is large, it is likely that many of them would be uninformative or redundant, therefore, initialization using such set might not provide a good starting point for the search.

## 5 Simulated data experiments

In this section we evaluate the proposed modeling framework for variable selection and adaptive classification through different simulated data experiments under various conditions. The objective is to assess the classification performance of the method, its ability of detecting the novel classes and its ability of discarding irrelevant variables and selecting those useful for classification.

### 5.1 Simulation study 1

This simulation study shows the usefulness of using all the variables available in the test data for class prediction and detection when only a small subset of these are observed in the training stage.

We consider the well known Italian wines dataset (Forina et al. 1986). The data consist of 27 chemical measurements from a collection of wine samples from Piedmont region, in Italy. The observations are classified into three classes indicating the type of wine. Different scenarios are considered for different combinations of number of variables observed in the training stage and different test data sample sizes. Using the class-specific sample means and covariances, we generate training data sets with random subsets of the 27 variables, with the number of variables observed in the training set $P$ equal to 18, 9, and 3. Then, with the same class-specific parameters, a test set on all the 27 variables and different sample sizes $N$ is generated. One class is randomly deleted from the training data, while all 3 classes are present in the test data. In each scenario, we consider the following models: the EDDA classifier fitted on the training data with full information, i.e. all 3 classes and all 27 variables, tested on the full test data, EDDA-full; the EDDA classifier fitted on the training data considering only a subset of the variables, then tested on the test data containing the same subset of training variables, EDDA; the AMDA approach of Bouveyron (2014) fitted on the simulated training data with a subset of the variables and tested on the test data with the subset of variables observed in the training, AMDA; the presented D-AMDA framework, D-AMDA. Further details are provided in Appendix 1.

Results are reported in Figs. 2 and 3. The variables in the wine data present a good degree of discrimination, and the EDDA model fitted and tested on the complete data represents the optimal baseline performance (EDDA-full). On the other hand,

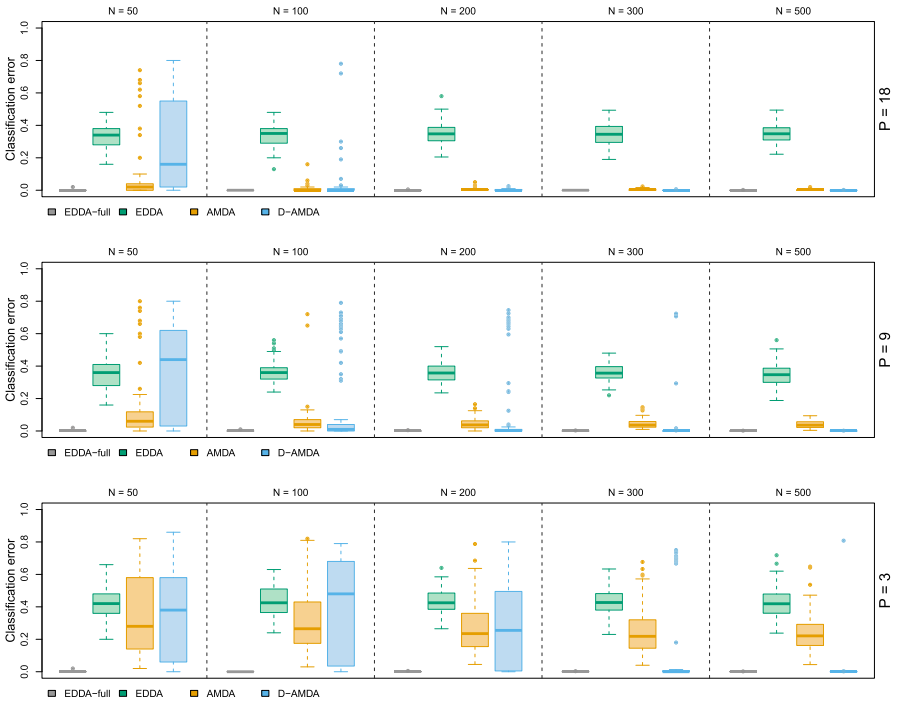



**Fig. 2** Simulation study 1 (Wine data). Classification error computed on the matched classes between the actual classification of the test data and the estimated one. The values are reported for different number of variables in the training stage $P$ and test data sample sizes $N$

the EDDA classifier trained on the partial data cannot account for the unobserved class in the test data and provides the worst classification performance. AMDA can detect additional classes in the test data, but it cannot use the discriminant information potentially available in the additional variables, thus obtaining an inferior classification performance compared to D-AMDA. Since the D-AMDA framework adapts to the additional dimensions and classes, all the information available in the variables observed in the test set is exploited for classification, of both variables observed during the training stage and the extra ones present in the test set. This extra information is beneficial, especially when the number of variables present in the training set is small ($P = 3$ in particular), attaining a classification performance comparable to the optimal baseline.

### 5.2 Simulation study 2

This simulation study assesses the D-AMDA classification performance and the effectiveness of the inductive variable selection method at detecting variables relevant for classification. Different scenarios are constructed by defining different proportions of relevant and irrelevant variables available in the training and the full test data.

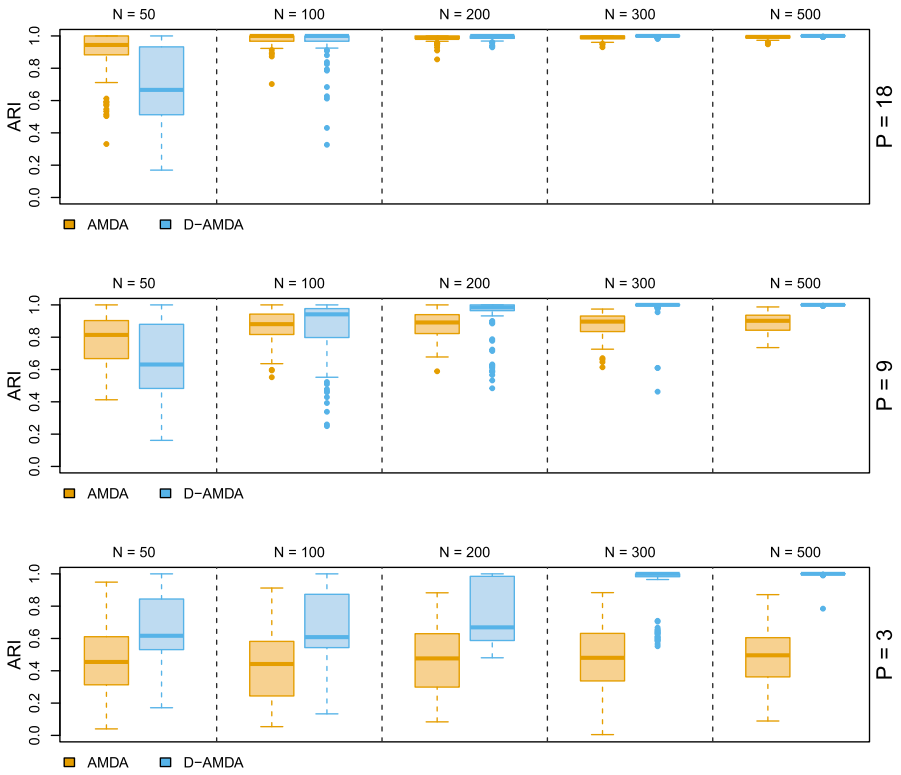



**Fig. 3** Simulation study 1 (Wine data). Adjusted Rand index between the actual classification of the test data and the estimated one for AMDA and D-AMDA. The values are reported for different number of variables in the training stage $P$ and test data sample sizes $N$

In all the experiments of this section we consider three types of variables: class-generative variables, Gen, which contain the principal information about the classes; redundant variables, Cor, which are correlated to the generative ones; noise variables Noi, which do not convey any information about the classes. The Gen variables are distributed according to a mixture of $C = 4$ multivariate Gaussian distributions. Each Cor variable is correlated to 2 Gen variables selected at random, while the Noi variables are independent from both Gen and Cor variables. We point out the fact that, as they are generated, the Cor variables actually contain some information about the classification. Indeed, they are independent of the label variable only conditionally on the set Gen, not marginally. Thus, in some cases, they could convey the best information available to classify the data units if some generative variables have been discarded during the search. Hence, the inclusion of a Cor variable would not necessarily degenerate the classification performance. In the learning set, 2 of the 4 classes are observed and they are randomly chosen. All the 4 classes are observed in the test set.

Three experiments are considered, each one characterized by three scenarios. Throughout the different scenarios, since D-AMDA is partially unsupervised, we use

the adjusted Rand index (ARI, Hubert and Arabie 1985) to assess the quality of the classification. We compare the results of the following methods:

- D-AMDA, the D-AMDA model applied on $\mathbf{X}$ and the full $\mathbf{Y}$ without performing any variable selection.
- D-AMDA-gen, the D-AMDA model applied on the learning and test sets containing only Gen variables. As only Gen variables are used, this represents the optimal baseline solution in terms of classification performance.
- D-AMDA-varsel, the D-AMDA model with the forward variable selection applied to the observed $\mathbf{X}$ and $\mathbf{Y}$.

The variable selection performance of D-AMDA-varsel is assessed via the proportion of times each variable was selected as relevant out of the total number of replicated experiments. Further details about the parameters of the simulations are in Appendix 1.

To evaluate the computational efficiency of D-AMDA with variable selection, we also report the computing times of D-AMDA-varsel in Appendix 1. All the experiments are run on a standard machine with 8 processors, implementing the variable selection search in parallel. The inductive framework for estimation and variable selection coupled with the parallelization of the forward greedy search is particularly efficient, with computing times having median values around the range of 20 to 150 seconds across all scenarios. More details are reported in Appendix 1.

#### 5.2.1 Experiment 1

The test data consist of 100 variables, 10 Gen, 30 Cor and 60 Noi. In the learning set, 20 of the 100 variables are observed. Three scenarios are defined according to the set of variables observed in the simulated $\mathbf{X}$:

1(a) All the 10 Gen variables plus 10 variables picked at random among Cor and Noi.
1(b) 5 Gen variables selected at random, 5 Cor selected at random, plus 10 variables chosen at random among Cor and Noi.
1(c) 2 Gen selected at random, the remaining 18 variables are chosen at random among Cor and Noi.

The sample size of the learning set is equal to the sample size of $\mathbf{Y}$ and takes values 100, 200 and 400. In all scenarios, the forward search is initialized starting from all the variables observed in $\mathbf{X}$.

Figures 4 and 5 report the results. In scenario 1(a), the EDDA learning model is estimated on a set containing all the classification variables. Furthermore, the forward search is initialized on the same set. This gives a good starting point to the variable selection procedure, resulting that only Gen variables are declared as relevant and with an excellent classification performance. The results hold regardless of the size of the test data samples in practice. In scenarios 1(b) and 1(c), as less Gen variables are available in the learning phase, the variable selection method declares as relevant Cor variables more frequently. However, good classification results and good selection performance are still obtained, especially for larger sample sizes.

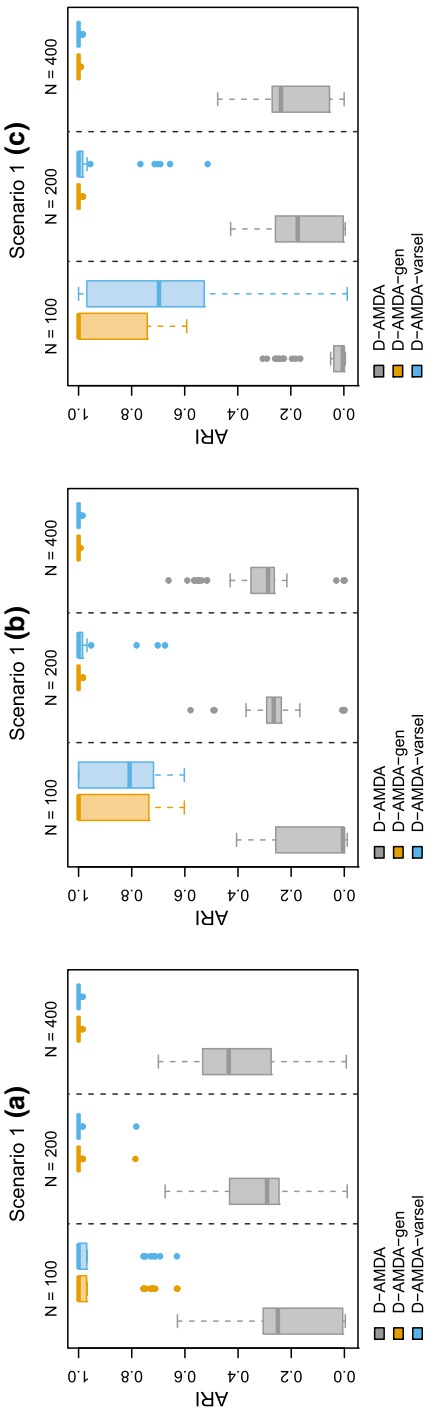



**Fig. 4** Simulation study 2, scenarios 1(a), 1(b), and 1(c). Adjusted Rand index between the actual classification of the test data and the estimated one for *D-AMDA*, *D-AMDA-gen*, and *D-AMDA-varsel*. The values are reported for different test data sample sizes $N$

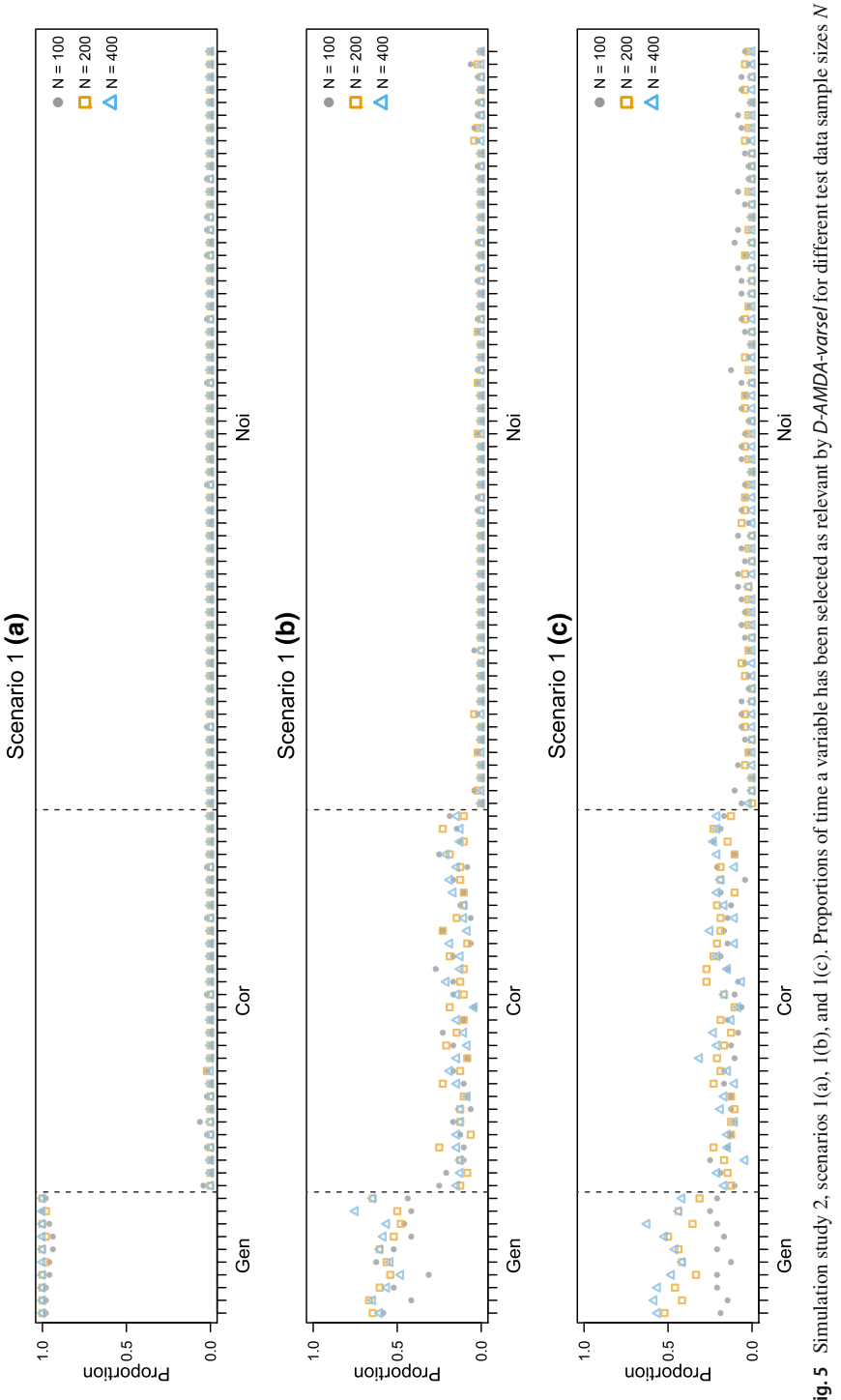



**Fig. 5** Simulation study 2, scenarios 1(a), 1(b), and 1(c). Proportions of time a variable has been selected as relevant by *D-AMDA-varsel* for different test data sample sizes *N*

### 5.2.2 Experiment 2

Also here the test data consist of 100 variables, 10 Gen, 30 Cor and 60 Noi. In the learning set, 50 of the 100 variables are observed. Three scenarios are defined according to the set of variables observed in the simulated $\mathbf{X}$:

2(a) All the 10 Gen variables plus 40 variables randomly chosen among Cor and Noi.
2(b) 5 Gen variables selected at random, 15 Cor selected at random, plus 30 variables chosen at random among Cor and Noi.
2(c) 2 Gen selected at random, the remaining 48 variables are randomly selected among Cor and Noi.

In this experiment, the sample size of the learning set is fixed and equal to 50 for all the scenarios. The forward search is initialized from 10 of the 50 variables observed in $\mathbf{X}$, selected using the ranking procedure described in Sect. 4.

This setting is particularly challenging, since the learning set is high-dimensional in comparison to the number of data points. In practice, this results in a learning phase where only EDDA models with diagonal covariance matrices can be estimated. Even if all Gen variables are observed in $\mathbf{X}$, such subset of models are misspecified in relation to how the data is generated. This represents a difficult starting point for the D-AMDA model and the variable selection procedure. Indeed, with this experiment we want to test the robustness of the method against the misspecification of the model in the learning stage. Results are reported in Figs. 6 and 7. In scenario 2(a), a selection of reasonable quality is attained, while in scenarios 2(b) and 2(c) Cor variables are selected almost as often as Gen variables. Overall, in all three scenarios, Noi variables are never selected and the method achieves a good classification performance even when Cor variables are selected as relevant almost as many times as the variables of the Gen set. This fact is likely due to the variable selection initialization: this initialization strategy tends to start the selection from a set of good classification variables, and such set may contain both Gen and Cor variables.

### 5.2.3 Experiment 3

In this case the test data consist of 200 variables, 20 Gen, 60 Cor and 120 Noi. In the learning set, 40 of the 200 variables are observed. Three scenarios are defined according to the set of variables observed in the simulated $\mathbf{X}$:

3(a) All the 20 Gen variables plus 20 variables selected randomly among Cor and Noi.
3(b) 10 Gen variables selected at random, 10 Cor selected at random, plus 20 variables picked at random among Cor and Noi.
3(c) 4 Gen selected at random, the remaining 36 variables are randomly chosen among Cor and Noi.

Here, the sample size of the learning set is equal to the one of the test data and takes values 100, 200 and 400. The forward search is initialized from 10 of the 40

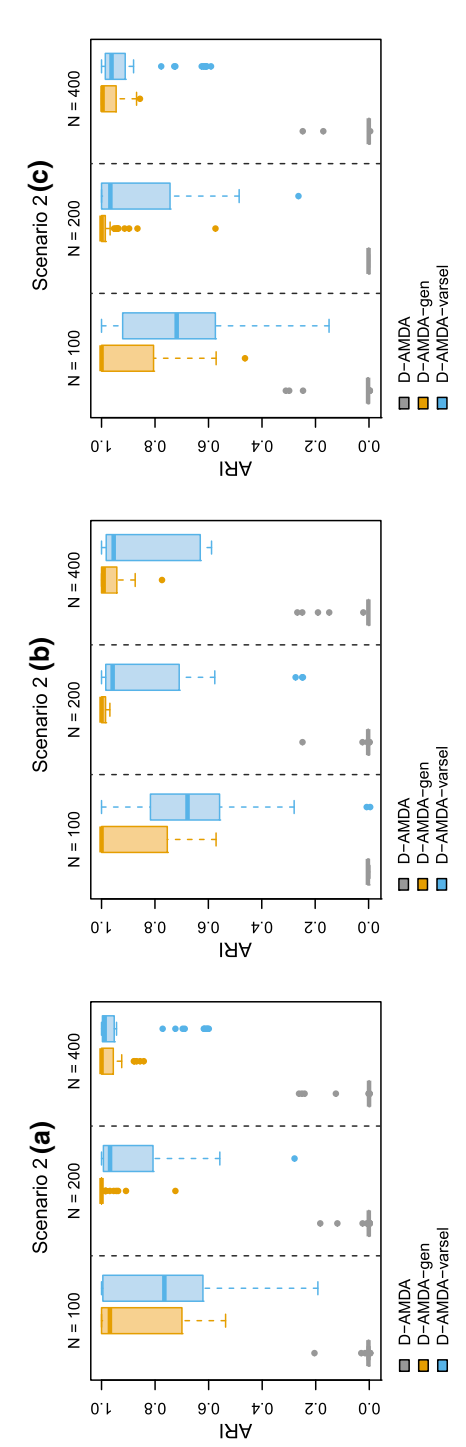



**Fig. 6** Simulation study 2, scenarios 2(a), 2(b), and 2(c). Adjusted Rand index between the actual classification of the test data and the estimated one for *D-AMDA*, *D-AMDA-gen*, and *D-AMDA-varsel*. The values are reported for different test data sample sizes $N$

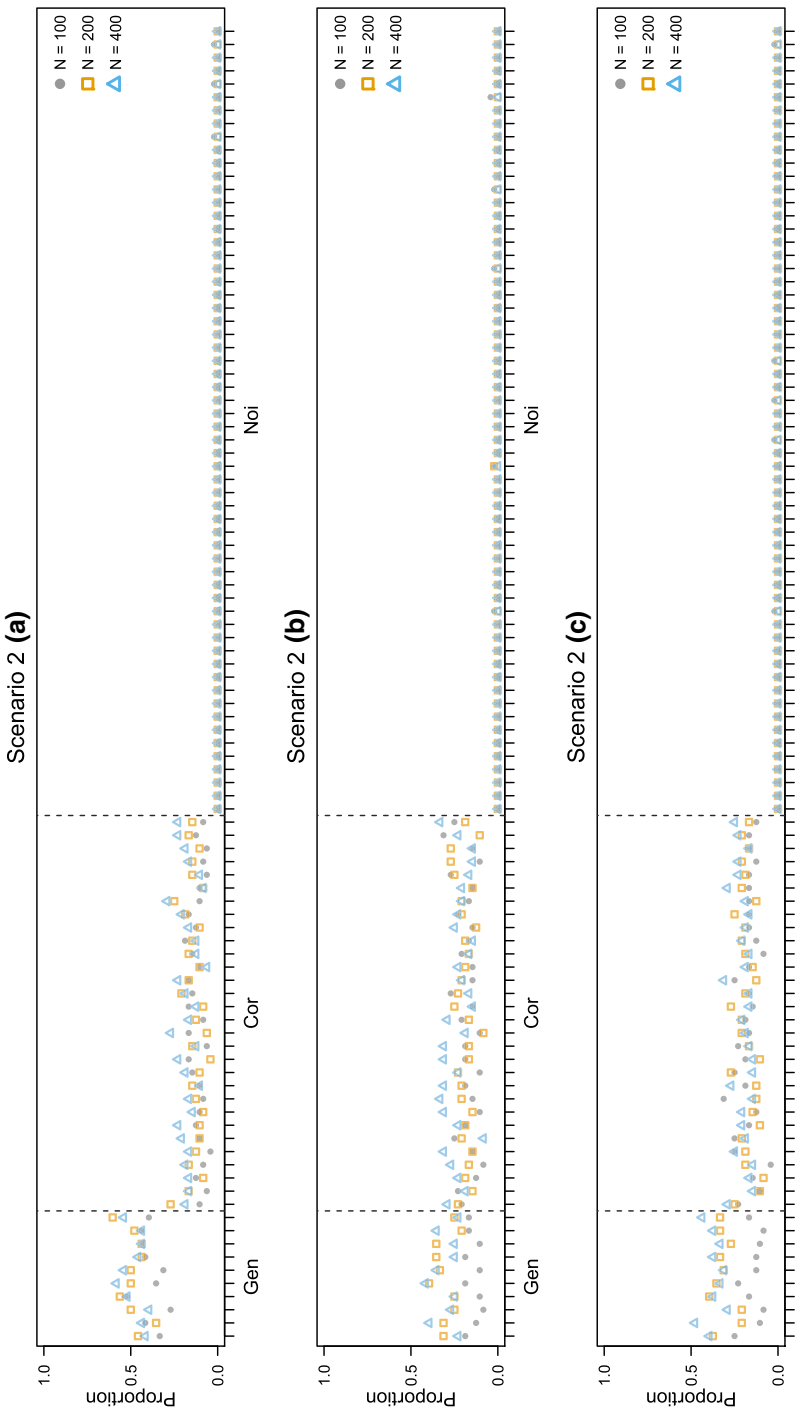



**Fig. 7** Simulation study 2, scenarios 2(a), 2(b), and 2(c). Proportions of time a variable has been selected as relevant by *D-AMDA-varsel* for different test data sample sizes *N*

variables observed in $\mathbf{X}$, selected using the ranking procedure described in Sect. 4. The experiment is characterized by an high-dimensional test set $\mathbf{Y}$.

Results are reported in Figs. 8 and 9. For larger sample sizes, the variable selection method tends to correctly identify the relevant variables, especially as the number of Gen variables involved in the estimation of the EDDA model in the learning phase increases, as in scenario 3(a). When such number is reduced in scenarios 3(b) and 3(c), Cor variables tend to be declared as relevant more often. However, Noi variables are never selected and the selection performance is still of sufficient quality for larger sample sizes. The D-AMDA method with variable selection obtains good classification results in all the scenarios.

## 6 Contaminated honey data

Food authenticity studies are concerned with establishing whether foods are authentic or not. Mid-infrared spectroscopy provides an efficient method of collecting data for use in food authenticity studies, without destructing the sample being tested nor requiring complex preparation (Downey 1996). In this section we consider a food authenticity data set consisting of mid-infrared spectroscopic measurements of honey samples. Kelly et al. (2006) collected 1090 absorbance spectra of artisanal Irish honey over the wavelength range $3700nm - 13600nm$ at $35nm$ resolution. Therefore, the data consists of 285 absorbance values (variables). Of these samples, 290 are pure honey, while the remaining are contaminated with five sugar syrups: beet sucrose (120), dextrose syrup (120), partial invert cane syrup (160), fully inverted beet syrup (280) and high-fructose corn syrup (120). The aim is to discriminate the pure honey from the adulterated samples and the different contaminants. At the same time, the purpose is in the identification of a small subset of absorbance values containing as much information for authentication purposes as the whole spectrum does. Figure 10 provides a graphical description of the data. Except from beet sucrose and dextrose, there is an high overlap between the other contaminants and the pure honey; this stems from the similar composition of honey and these syrups (Kelly et al. 2006). The principal features seem to be around the ranges 8700 nm–10,300 nm and 10,500 nm–11,600 nm, while the spectra overlap significantly at lower wavelengths.

In this section we test the D-AMDA method with variable selection. We construct an artificial experiment that represents the situation were the samples in the learning set were collected at a lower resolution than the ones in test data and the information about one of the contaminants was missing. We randomly split the whole data into learning set and test set, in proportions $2/3$ and $1/3$ respectively. Then, we consider the learning set as it were generated from absorbance spectra collected at 70 nm intervals, retaining wavelengths 3700 nm, 3770 nm, 3840 nm and so on. Thus, the data observed in the learning phase are approximately recorded on half of the variables of the test data. Afterwards, we randomly chose one of the two classes related to the contaminants beet sucrose and dextrose syrup, and we remove from the learning set the corresponding observations. In this way we obtain a test set measured on additional variables and containing extra classes.

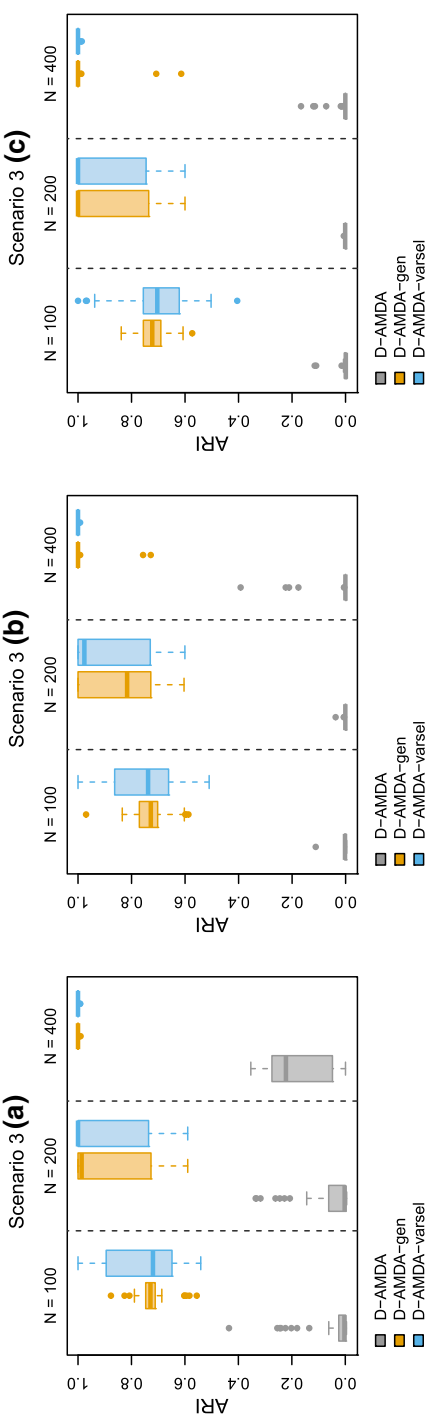



**Fig. 8** Simulation study 2, scenarios 3(a), 3(b), and 3(c). Adjusted Rand index between the actual classification of the test data and the estimated one for *D-AMDA*, *D-AMDA-gen*, and *D-AMDA-varsel*. The values are reported for different test data sample sizes $N$

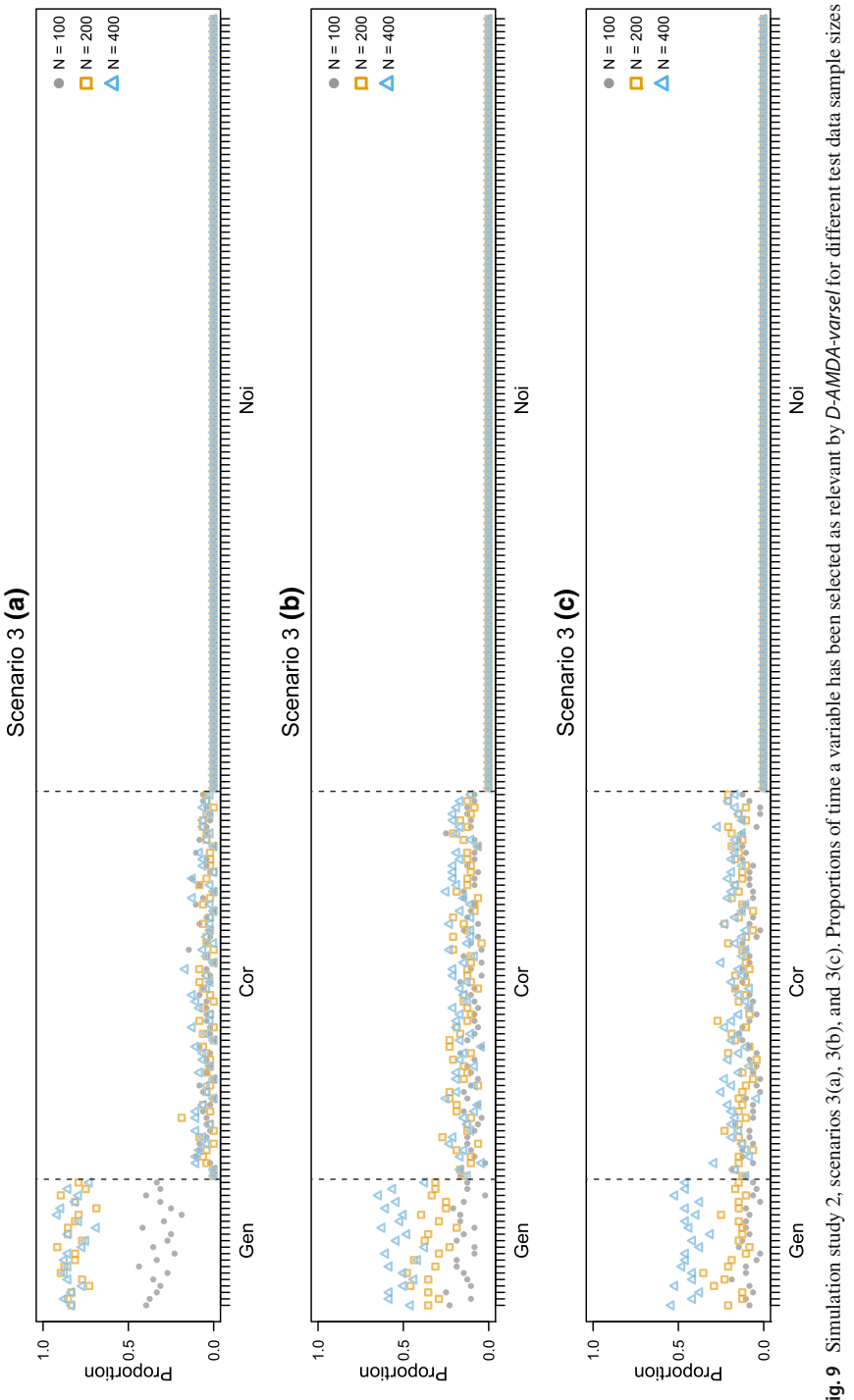



**Fig. 9** Simulation study 2, scenarios 3(a), 3(b), and 3(c). Proportions of time a variable has been selected as relevant by *D-AMDA-varsel* for different test data sample sizes *N*

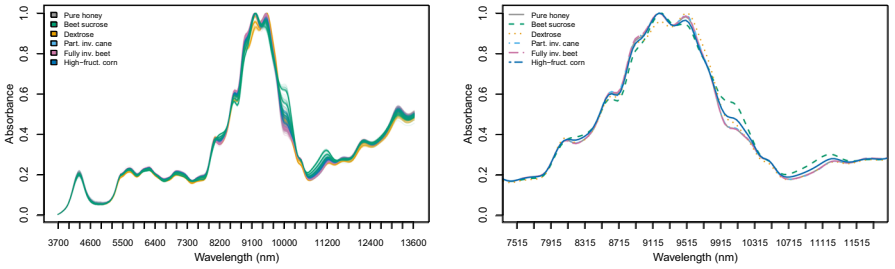



**Fig. 10** The Mid-infrared spectra recorded for the pure and the contaminated honey samples (**a**); class-conditional mean spectra for pure and contaminated honey samples, zoomed around the range 7500 nm–11,700 nm (**b**)

We replicate the experiment for 100 times, applying the D-AMDA approach with variable selection, D-AMDA-varsel, and without, D-AMDA. For comparison and for evaluating the classification performance, we also apply the EDDA model to the whole learning set containing all the contaminants, EDDA-full. Then we use the estimated classifier on the test data to classify the samples. The EDDA-full classifier uses all the information available about classes and wavelengths, thus its classification performance can be considered as the optimal baseline. For the variable selection, we initialize the search from a set of 30 wavelengths selected using the ranking procedure described in Sect. 4.

Results of the variable selection procedure are reported in Fig. 11. The figure displays the proportion of times a wavelength has been declared as relevant for separating the classes of contaminants and the pure honey. The frequently chosen wavelengths are mostly in the ranges 10,000 nm–11,200 nm and 8500 nm–9300 nm. In particular, values 10,000 nm, 10,070 nm, 10,140 nm, 10,210 nm, 10,910 nm, 10,980 nm, 11,050 nm, and 11,120 nm are selected in all the replicates of the experiment. Also wavelengths in the range 5400 nm–5800 nm are selected a significant number of times. The peak in 8250 nm corresponds to a wavelength range particularly useful to discriminate dextrose syrup from the rest (Kelly et al. 2006). The most frequently selected wavelengths correspond to the interesting peaks and features of the spectra. Classification results are presented in Figure 12. As for the simulation settings, because of the extra hidden class in the test set, we made use of the ARI to compare the actual classification and the ones estimated by D-AMDA-varsel and D-AMDA. The D-AMDA method selects the correct number of classes only 34/100 of the times. D-AMDA-varsel selects the right number of unknown classes 79 out of 100 times, and panel (b) of Figure 12 reports the boxplot of the classification error of D-AMDA-varsel and EDDA-full in this case. The classification performance of D-AMDA-varsel is comparable to EDDA-full, but it makes use of information about less wavelengths and is obtained in a more complex setting.

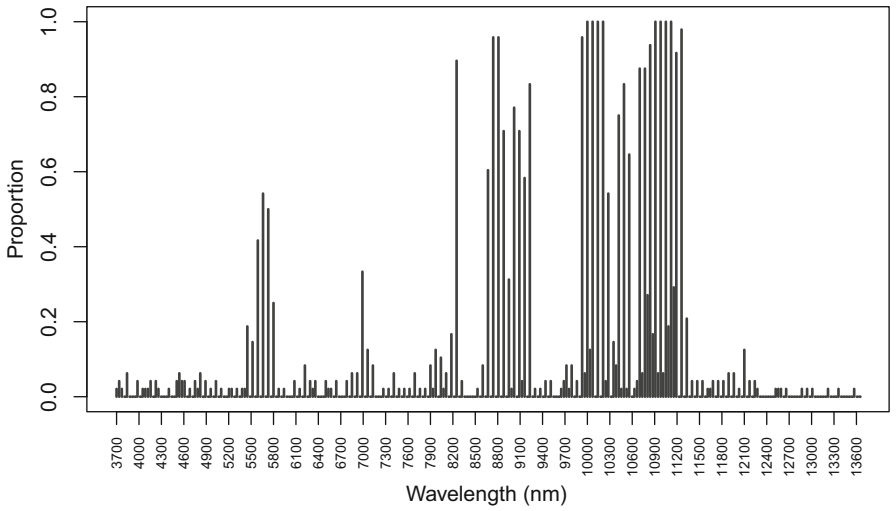



**Fig. 11** Proportions of time a wavelength has been selected as a relevant variable over 100 replicates of the artificial experiment

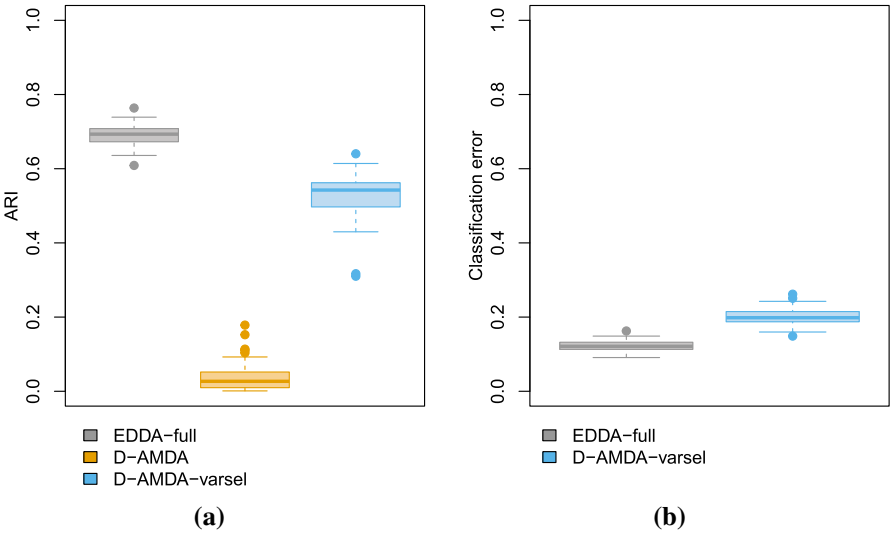



**Fig. 12** ARI **(a)** and classification error **(b)**. The classification error is reported for the EDDA model and the D-AMDA with variable selection. For the error, the boxplot displays values only for the 79/100 times the D-AMDA correctly selected the number of unknown classes

## 7 Discussion

We presented a general adaptive mixture discriminant analysis method for classification and variable selection when the test data contain unobserved classes and extra variables. We have shown that our methodology effectively addresses the issues gen-

erated by the presence of hidden classes in a test data with augmented dimensions compared to the data observed during the training stage. As such, the method is suitable for applications in real-time classification problems where the new data points to be labelled convey extra information thanks to the presence of additional input features.

The inductive approach had the advantage of avoiding the storing of the learning set and of avoiding the re-estimation of the parameters already obtained in the learning stage. However, when extra variables are observed in the test data, the estimation process is a complex problem, due to the parameter constraints induced by the initial learning phase. An inductive conditional estimation procedure has been introduced to overcome the issue and obtain valid parameter estimates related to the added dimensions. The inductive framework results in a fast and computationally efficient procedure, which has been embedded into a variable selection method for dealing with high-dimensional data.

The D-AMDA method developed here lies within the framework of novelty detection and dataset shift. Novelty detection is the identification of unknown classes that a classification system is not aware of during training (Markou and Singh 2003), while more in general dataset shift refers to the difference between the joint distribution of labels and input variables in the training and test sets (Quionero-Candela et al. 2009; Moreno-Torres et al. 2012). Compared to the AMDA method of Bouveyron (2014), in addition to the problem of unrepresented classes in the training set, in this paper we also address the shift due to the increased dimensions of the test data, which leads to a change in the distributions of input variables in training and test sets through the different sizes of the parameter spaces. The problem of test and training data having different dimensions could be viewed as an instance of a particular type of dataset shift, linked to "covariate shif" (see Moreno-Torres et al. 2012, for more details). To the best of our knowledge, this problem has only been scantily explored in the literature, often with a focus to specific related applications. For example, particular cases of classification of training and test data with different dimensions are those of supervised classification of time series of varying lengths (Tan et al. 2019; Bagnall et al. 2017, for recent reviews) and classification of images with occlusions and corruptions (see for example Zhou et al. 2009; Bao et al. 2013, for an overview of the problem). The D-AMDA method presented here provides scope for future developments to deal with these specific complex situations.

The proposed D-AMDA framework opens also interesting future methodological research directions. A limitation of the D-AMDA approach is that the discovery phase does not consider particular constraints on the estimated covariance matrices. The introduction of parsimonious models as in Bensmail and Celeux (1996) and Cappozzo et al. (2020) with adaptive dimensions may be object of future research. Another limitation is that the labels observed in the training data are assumed to be noise free, as well as that no outlier observations are present in the input features. Recent work by Cappozzo et al. (2020) proposes a robust version of the AMDA framework to address these added sources of complexity. Future work may explore the development of a robust version of D-AMDA, with a particular focus on discarding those additional dimensions characterized by high levels of noisy and contaminated observations, suitable for robust on-line classification.

## 8 Software

The R package `damda` implements the D-AMDA framework with inductive variable selection presented in this paper. The package is publicly available at one of the authors webpage: https://michaelfop.github.io/.

**Acknowledgements** The authors would like to thank the editor and the anonymous referees for their valuable comments, which helped in substantially improving the quality of this paper. The work of Fop M. and Murphy T. B. was supported by the Science Foundation Ireland funded Insight Research Centre (SFI/12/RC/2289_P2). The work of Mattei P. A. and Bouveyron C. has been supported by the French government, through the 3IA Côte d'Azur Investment in the Future project managed by the National Research Agency (ANR) with the reference numbers ANR-19-P3IA-0002.



## Appendix

### A. Details of the inductive conditional estimation

Neglecting the term involving the mixing proportions, the objective function to be optimized in the M step to estimate $\boldsymbol{\mu}_k^*$ and $\boldsymbol{\Sigma}_k^*$ is given by:

$$F(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) = \sum_{i=1}^{N} \left[ \sum_{k=1}^{K} t_{ik} \log \left\{ \phi(\mathbf{y}_i ; \boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k^*) \right\} \right].$$

Let $\mathbf{O}_k = \sum_{i=1}^{N} t_{ik} (\mathbf{y}_i - \bar{\mathbf{y}}_k)(\mathbf{y}_i - \bar{\mathbf{y}}_k)'$, with $\bar{\mathbf{y}}_k = \frac{1}{N_k} \sum_{i=1}^{N} t_{ik}\, \mathbf{y}_i$. The above function can be expressed in term of the covariance matrix as:

$$F(\boldsymbol{\Sigma}^*) = \sum_{k} \operatorname{tr}\left\{ \mathbf{O}_k (\boldsymbol{\Sigma}_k^*)^{-1} \right\} + \sum_{k} N_k \log \det \boldsymbol{\Sigma}_k^*.$$

Let us now consider the partitioned matrices:

$$\boldsymbol{\Sigma}_k^* = \begin{bmatrix} \bar{\boldsymbol{\Sigma}}_k & \mathbf{C}_k \\ \mathbf{C}_k' & \boldsymbol{\Sigma}_k^{\mathrm{Q}} \end{bmatrix}, \qquad \mathbf{O}_k = \begin{bmatrix} \mathbf{W}_k & \mathbf{V}_k \\ \mathbf{V}_k' & \mathbf{U}_k \end{bmatrix}.$$

Furthermore, define $\mathbf{E}_k = \boldsymbol{\Sigma}_k^{\mathrm{Q}} - \mathbf{C}_k' \bar{\boldsymbol{\Sigma}}_k^{-1} \mathbf{C}_k$. Then

$$(\boldsymbol{\Sigma}_k^*)^{-1} = \begin{bmatrix} \bar{\boldsymbol{\Sigma}}_k^{-1} + \bar{\boldsymbol{\Sigma}}_k^{-1} \mathbf{C}_k \mathbf{E}_k^{-1} \mathbf{C}_k' \bar{\boldsymbol{\Sigma}}_k^{-1} & -\bar{\boldsymbol{\Sigma}}_k^{-1} \mathbf{C}_k \mathbf{E}_k^{-1} \\ \mathbf{E}_k^{-1} \mathbf{C}_k' \bar{\boldsymbol{\Sigma}}_k^{-1} & \mathbf{E}_k^{-1} \end{bmatrix},$$

and $\log\det\boldsymbol{\Sigma}_k^* = \log\det\overline{\boldsymbol{\Sigma}}_k + \log\det\mathbf{E}_k$. It follows that $F(\boldsymbol{\Sigma}^*)$ can be re-expressed as function of $\mathbf{E}_k$ and $\mathbf{C}_k$ as follows:

$$F(\mathbf{E},\mathbf{C}) = \sum_k \mathrm{tr}\{\mathbf{W}_k\overline{\boldsymbol{\Sigma}}_k^{-1}\mathbf{C}_k\mathbf{E}_k^{-1}\mathbf{C}_k'\overline{\boldsymbol{\Sigma}}_k^{-1}\} - 2\sum_k \mathrm{tr}\{\mathbf{V}_k'\overline{\boldsymbol{\Sigma}}_k^{-1}\mathbf{C}_k\mathbf{E}_k^{-1}\} + \sum_k \mathrm{tr}\{\mathbf{U}_k\mathbf{E}_k^{-1}\} + \sum_k N_k \log\det\mathbf{E}_k + \text{const.}$$

Maximization of $F(\mathbf{E},\mathbf{C})$ with respect to $\mathbf{E}_k$ and $\mathbf{C}_k$ leads to:

$$\widehat{\mathbf{C}}_k = (\overline{\boldsymbol{\Sigma}}_k^{-1}\mathbf{W}_k\overline{\boldsymbol{\Sigma}}_k^{-1})^{-1}(\overline{\boldsymbol{\Sigma}}_k^{-1}\mathbf{V}_k),$$
$$\widehat{\mathbf{E}}_k = \frac{1}{N_k}\left[\widehat{\mathbf{C}}_k'\overline{\boldsymbol{\Sigma}}_k^{-1}\mathbf{W}_k\overline{\boldsymbol{\Sigma}}_k^{-1}\widehat{\mathbf{C}}_k - 2\mathbf{V}_k'\overline{\boldsymbol{\Sigma}}_k^{-1}\widehat{\mathbf{C}}_k + \mathbf{U}_k\right].$$

Consequently we have that:

$$\widehat{\boldsymbol{\Sigma}}_k^{\mathrm{Q}} = \widehat{\mathbf{E}}_k + \widehat{\mathbf{C}}_k'\overline{\boldsymbol{\Sigma}}_k^{-1}\widehat{\mathbf{C}}_k.$$

Given estimates $\widehat{\mathbf{C}}_k$ and $\widehat{\mathbf{E}}_k$, for the mean parameter $\boldsymbol{\mu}_k^{\mathrm{Q}}$ corresponding to the additional variables, define now $\mathbf{m}_{ik} = \boldsymbol{\mu}_k^{\mathrm{Q}} + \mathbf{C}_k'\overline{\boldsymbol{\Sigma}}_k^{-1}(\mathbf{y}_i^{\mathrm{P}} - \overline{\boldsymbol{\mu}}_k)$. Consequently, the function $F(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ can be rewritten as:

$$F(\mathbf{m}) = \sum_{i=1}^{N}\left[\sum_{k=1}^{K} t_{ik}\log\left\{\phi(\mathbf{y}_i^{\mathrm{Q}} \mid \mathbf{y}_i^{\mathrm{P}};\ \mathbf{m}_{ik}, \widehat{\mathbf{E}}_k)\right\}\right] + \text{const.}$$

By plugging the $\mathbf{m}_{ik}$ expression above in $F(\mathbf{m})$, we can express the latter in terms of $\boldsymbol{\mu}_k^{\mathrm{Q}}$ as:

$$F(\boldsymbol{\mu}^{\mathrm{Q}}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}\left\{\left[\mathbf{y}_i^{\mathrm{Q}} - \boldsymbol{\mu}_k^{\mathrm{Q}} - \widehat{\mathbf{C}}_k'\overline{\boldsymbol{\Sigma}}_k^{-1}(\mathbf{y}_i^{\mathrm{P}} - \overline{\boldsymbol{\mu}}_k)\right]'\widehat{\mathbf{E}}_k^{-1}\left[\mathbf{y}_i^{Q} - \boldsymbol{\mu}_k^{\mathrm{Q}} - \widehat{\mathbf{C}}_k'\overline{\boldsymbol{\Sigma}}_k^{-1}(\mathbf{y}_i^{\mathrm{P}} - \overline{\boldsymbol{\mu}}_k)\right]\right\} + \text{const.}$$

Taking derivatives of $F(\boldsymbol{\mu}^{\mathrm{Q}})$ and solving for $\boldsymbol{\mu}_k^{\mathrm{Q}}$ we obtain:

$$\widehat{\boldsymbol{\mu}}_k^{\mathrm{Q}} = \frac{1}{N_k}\left[\sum_{i=1}^{N} t_{ik}\mathbf{y}_i^{\mathrm{Q}} - \widehat{\mathbf{C}}_k'\overline{\boldsymbol{\Sigma}}_k^{-1}\sum_{i=1}^{N} t_{ik}(\mathbf{y}_i^{\mathrm{P}} - \overline{\boldsymbol{\mu}}_k)\right].$$

The above passages prove the derivation of the updating equations of the M step in Sect. 3.2.2.

## B. A note on regularization

The procedure described in 3.2.2 requires the empirical class scatter matrix $\mathbf{O}_k$ to be definite positive. This may not be the case in situations where the expected number of observations in a class is small or the variables are highly correlated. Approaches for Bayesian regularization in the context of finite Gaussian mixture models for clustering have already been suggested in the literature, see in particular Baudry and Celeux (2015). We suggest a similar approach, proposing the following regularized version of $\mathbf{O}_k$:

$$\mathbf{O}_k^{\text{reg}} = \mathbf{O}_k + \frac{\mathbf{S}}{N \det(\mathbf{S})^{1/R}} \left( \frac{\gamma}{K + H} \right)^{1/R},$$

where $\mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'$ is the empirical covariance matrix computed on the full test data, and $\bar{\mathbf{y}}$ the sample mean, $\bar{\mathbf{y}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{y}_i$. The second term of the sum is a matrix whose determinant is proportional to $\gamma/(K + H)$ and acts as a regularizer. The coefficient $\gamma$ controls the amount of regularization and we set it to $(\log R)/N^2$; see Baudry and Celeux (2015) for further details. Note that in the case where $N \leq R$, the sample covariance matrix $\mathbf{S}$ is replaced by the diagonal matrix $\text{diag}(\mathbf{S})$.

## C. Details of simulation experiments

In this section we describe in more details the settings of the simulated data experiments of Sect. 5 in the main text.

### C.1. Simulation study 1

The training data has $M = 300$ observations in all scenarios. A random subset of the 27 variables is taken from the data, with the number of variables observed in the training set equal to $P = \{18, 9, 3\}$. The test set is generated using all the 27 variables, considering different sample sizes $N = \{50, 100, 200, 300, 500\}$. Different scenarios are defined by different combinations of $P$ and $N$. One class is randomly deleted from the training data, while all 3 classes are present in the test data. In each scenario, the following models are considered: EDDA-full, the EDDA classifier fitted on the training data with full information, i.e. all 3 classes and all 27 variables, tested on the full test data; EDDA the EDDA classifier fitted on the training data considering only a subset of the variables, then tested on the full test data; the AMDA approach of Bouveyron (2014) fitted on the simulated training data with a subset of the variables and tested on the test data with the subset of variables observed in the training; the presented D-AMDA framework. Each experiment is replicated 100 times for all combinations of sample sizes and number of observed training variables. Model selection for AMDA and D-AMDA is performed using BIC and a range of values of $H$ from 0 to 4. Since AMDA and D-AMDA are partially unsupervised, we compute the classification error on the matching classes detected after tabulating the actual classification with the estimated

one using function `matchClasses` of package `e1071` (Meyer et al. 2019). To compare AMDA and D-AMDA, we also report the adjusted Rand index (ARI, Hubert and Arabie 1985). Indeed, the learning in the test set is partly unsupervised, and a number of hidden classes different from 1 could be estimated. The results are reported in Figs. 2 and 3 in the main text.

### C.2. Simulation study 2

The data are generated according to the following settings and parameters. Gen variables are distributed according to a mixture of $C = 4$ multivariate Gaussian distributions with mixing proportions (0.3, 0.4, 0.4, 0.3). Mean parameters are randomly chosen in $(-7, 7)$, $(-4.5, 4.5)$, $(-0.5, 0.5)$, $(-10, 10)$. For each class, the covariance matrices are randomly generated from the Wishart distributions $\mathcal{W}(G, \Psi_1)$, $\mathcal{W}(G+2, \Psi_2)$, $\mathcal{W}(G+1, \Psi_3)$, $\mathcal{W}(G, \Psi_4)$, where $G$ denotes the number of generative variables. The scale matrices are respectively defined: $\Psi_1$, is such that $\psi_{jj} = 1$ and $\psi_{ji} = \psi_{ij} = 0.7$; $\Psi_3$, is such that $\psi_{jj} = 1$ and $\psi_{ji} = \psi_{ij} = 0.5$; $\Psi_2 = \Psi_4 = \mathbb{I}$. Cor variables are generated as $X_{g_1} + X_{g_2} + \epsilon$, where $X_{g_1}$ and $X_{g_2}$ are two randomly chosen Gen variables and $\epsilon \sim \mathcal{N}(0, 1)$. In Simulations 1 and 2, Noi variables are generated as $\mathcal{N}(\mathbf{0}, \Psi)$, where $\Psi$ is such that $\psi_{jj} = 1$ and $\psi_{ji} = \psi_{ij} = 0.5$; thus they are correlated to each other, but not to Cor and Gen variables. In Experiment 3, the Noi variables are generated all independent of each other. The 2 classes observed in the learning set are randomly chosen from the set of 4 classes with equal probabilities.

We considered three different sample sizes for the test data, respectively 100, 200 and 400. Each scenario within each experiment and for each sample size was replicated 50 times. Throughout the different scenarios, we compared the results of the following methods: D-AMDA, the D-AMDA model applied on $\mathbf{X}$ and the full $\mathbf{Y}$ without performing any variable selection; D-AMDA-gen, representing the optimal baseline solution, which corresponds to the D-AMDA model applied on the learning and test sets where only Gen variables are observed; D-AMDA-varsel, the D-AMDA model with the forward variable selection applied to the observed $\mathbf{X}$ and $\mathbf{Y}$.

We used the ARI to assess the quality of the classification of all methods, while the variable selection performance of D-AMDA-varsel was assessed via the proportion of times each variable was selected as relevant out of the 50 replicated experiments. The results are reported in the Figs. 4, 5, 6, 7, 8, and 9 in the main text.

## D. Computing times

To evaluate the computational efficiency of D-AMDA with variable selection, we report the computing times of the Simulation Study 2 experiments of Sect. 5. All the experiments were run on a standard machine with 8 processors (a Dell laptop Intel® Core i7-8650U CPU @1.90GHz×8). The code implementing the proposed framework is mainly written in R, with parts of the estimation procedure written in C++; the greedy forward search is implemented using the standard parallelization functionalities of R. Figure 13 shows the computing times (in seconds) for all scenarios

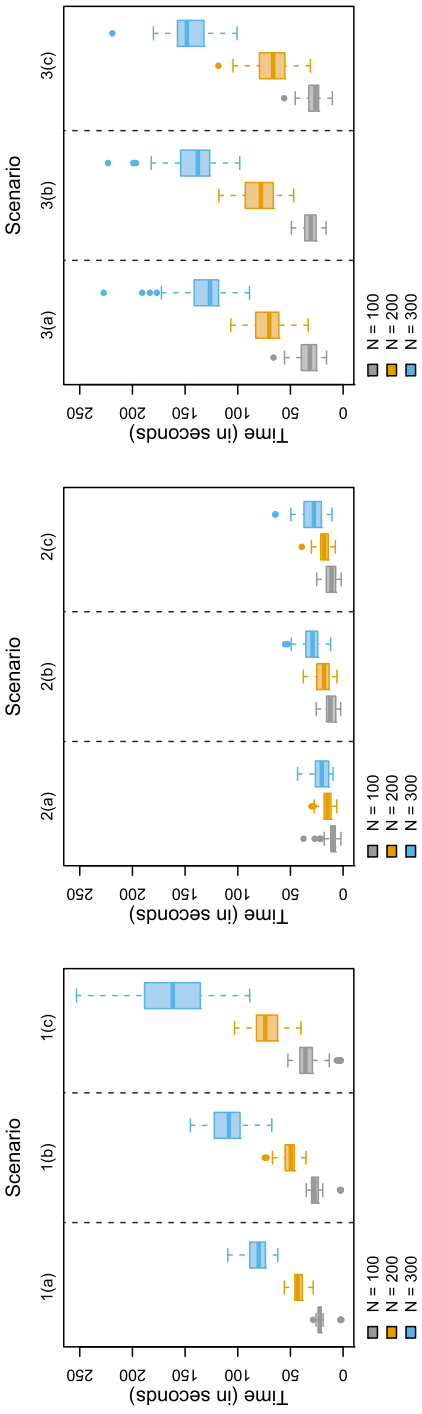



**Fig. 13** Simulation study 2. Computing times in seconds for D-AMDA with variable selection for the different scenarios and test data sample sizes

of Simulation Study 2. The inductive framework for estimation and variable selection coupled with the parallelization of the forward greedy search is particularly efficient, with computing times having median values around the range of 20 to 150 s across all scenarios.

We want to note that evaluating the effective runtime and speed of a method is a very difficult task; we point the interested reader to Kriegel et al. (2017) for a discussion.

## References


Bagnall A, Lines J, Bostrom A, Large J, Keogh E (2017) The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Min Knowl Disc 31(3):606–660

Bao B-K, Liu G, Hong R, Yan S, Xu C (2013) General subspace learning with corrupted training data via graph embedding. IEEE Trans Image Process 22(11):4380–4393

Baudry J-P, Celeux G (2015) EM for mixtures Initialization requires special care. Stat Comput 25(4):713–726

Bazell D, Miller DJ (2005) Class discovery in galaxy classification. Astrophys J 618(2):723

Bensmail H, Celeux G (1996) Regularized Gaussian discriminant analysis through eigenvalue decomposition. J Am Stat Assoc 91:1743–1748

Bouveyron C (2014) Adaptive mixture discriminant analysis for supervised learning with unobserved classes. J Classif 31(1):49–84

Bouveyron C, Celeux G, Murphy TB, Raftery AE (2019) Model-based clustering and classification for data science: with applications in R, vol 50. Cambridge University Press, Cambridge

Brusco MJ, Steinley D (2011) Exact and approximate algorithms for variable selection in linear discriminant analysis. Comput Stat Data Anal 55(1):123–131

Cappozzo A, Greselin F, Murphy TB (2020) Anomaly and novelty detection for robust semi-supervised learning. Stat Comput 30(5):1545–1571

Celeux G, Govaert G (1995) Gaussian parsimonious clustering models. Pattern Recogn 28(5):781–793

Chapelle O, Schölkopf B, Zien A (eds) (2006) Semi-Supervised learning. MIT Press

Chaudhuri S, Drton M, Richardson TS (2007) Estimation of a covariance matrix with zeros. Biometrika 94(1):199–216

Clemmensen L, Hastie T, Witten D, Ersbøll B (2011) Sparse discriminant analysis. Technometrics 53(4):406–413

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc B 39(1):1–38

Downey G (1996) Authentication of food and food ingredients by near infrared spectroscopy. J Near Infrared Spectrosc 4(1):47–61

Fienberg SE, Meyer MM (2006) Iterative proportional fitting. Encyclop Stat Sci 6:3723–3726

Fop M, Murphy TB (2018) Variable selection methods for model-based clustering. Stat Surv 12:18–65

Forina M, Armanino C, Castino M, Ubigli M (1986) Multivariate data analysis as a discriminating method of the origin of wines. Vitis 25(3):189–201

Fraley C (1998) Algorithms for model-based Gaussian hierarchical clustering. SIAM J Sci Comput 20(1):270–281

Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis and density estimation. J Am Stat Assoc 97:611–631

Frame SJ, Jammalamadaka SR (2007) Generalized mixture models, semi-supervised learning, and unknown class inference. Adv Data Anal Classif 1(1):23–38

Friedman JH (1989) Regularized discriminant analysis. J Am Stat Assoc 84(405):165–175

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182

Hastie T, Tibshirani R (1996) Discriminant analysis by Gaussian mixtures. J Royal Stat Soc Ser B (Methodological) 58(1):155–176

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2:193–218

Jiang B, Wang X, Leng C (2018) A direct approach for sparse quadratic discriminant analysis. J Mach Learn Res 19(1):1098–1134

Kelly JD, Petisco C, Downey G (2006) Application of fourier transform midinfrared spectroscopy to the discrimination between irish artisanal honey and such honey adulterated with various sugar syrups. J Agric Food Chem 54(17):6166–6171
Kriegel H-P, Schubert E, Zimek A (2017) The (black) art of runtime evaluation: Are we comparing algorithms or implementations? Knowl Inf Syst 52(2):341–378
Lawoko C, McLachlan G (1989) Bias associated with the discriminant analysis approach to the estimation of mixing proportions. Pattern Recogn 22(6):763–766
Le KT, Chaux C, Richard FJ, Guedj E (2020) An adapted linear discriminant analysis with variable selection for the classification in high-dimension, and an application to medical data. Comput Stat Data Anal 152:107031
Mai Q, Zou H, Yuan M (2012) A direct approach to sparse discriminant analysis in ultra-high dimensions. Biometrika 99(1):29–42
Markou M, Singh S (2003) Novelty detection: a review-part 1: statistical approaches. Signal Process 83(12):2481–2497
Maugis C, Celeux G, Martin-Magniette ML (2009a) Variable selection for clustering with Gaussian mixture models. Biometrics 65:701–709
Maugis C, Celeux G, Martin-Magniette ML (2009b) Variable selection in model-based clustering: a general variable role modeling. Comput Stat Data Anal 53:3872–3882
Maugis C, Celeux G, Martin-Magniette ML (2011) Variable selection in model-based discriminant analysis. J Multivar Anal 102(10):1374–1387
McLachlan G (2004) Discriminant analysis and statistical pattern recognition. Wiley, New York
McLachlan G, Krishnan T (2008) The EM algorithm and extensions. Wiley, New York
McLachlan GJ (2012) Discriminant analysis. Wiley Interdis Rev Comput Stat 4(5):421–431
Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2019) e1071: Misc functions of the department of statistics, probability theory group (Formerly: E1071), TU Wien. R package version 1.7-3
Miller DJ, Browning J (2003) A mixture model and EM-based algorithm for class discovery, robust classification, and outlier rejection in mixed labeled/unlabeled data sets. IEEE Trans Pattern Anal Mach Intell 25(11):1468–1483
Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F (2012) A unifying view on dataset shift in classification. Pattern Recogn 45(1):521–530
Murphy TB, Dean N, Raftery AE (2010) Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. Ann Appl Stat 4(1):396–421
Pacheco J, Núñez SC, Gómez O (2006) Analysis of new variable selection methods for discriminant analysis. Comput Stat Data Anal 51(3):1463–1478
Pang S, Kasabov N. (2004). Inductive vs transductive inference, global vs local models: SVM, TSVM, and SVMT for gene expression classification problems. In: 2004 IEEE international joint conference on neural networks, vol 2, pp 1197–1202
Qin Y (2018) A review of quadratic discriminant analysis for high-dimensional data. Computational Statistics, Wiley Interdisciplinary Reviews
Quionero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND (2009) Dataset shift in machine learning. The MIT Press, Cambridge
Raftery AE, Dean N (2006) Variable selection for model-based clustering. J Am Stat Assoc 101:168–178
Safo SE, Ahn J (2016) General sparse multi-class linear discriminant analysis. Comput Stat Data Anal 99:81–90
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464
Scrucca L, Fop M, Murphy TB, Raftery AE (2016) mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. The R J 8(1):289–317
Scrucca L, Raftery AE (2015) Improved initialisation of model-based clustering using Gaussian hierarchical partitions. Adv Data Anal Classif 9(4):447–460
Scrucca L, Raftery AE (2018) Clustvarsel: a package implementing variable selection for Gaussian model-based clustering in R. J Stat Softw Articles 84(1):1–28
Sun J, Zhao H (2015) The application of sparse estimation of covariance matrix to quadratic discriminant analysis. BMC Bioinformatics 16(1)
Tan C. W, Petitjean F, Keogh E, Webb G. I. (2019). Time series classification for varying length series. arXiv:1910.04341
Tong Y (1990) The multivariate normal distribution. Springer, Berlin
Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley, New York

Woillez M, Ressler PH, Wilson CD, Horne JK (2012) Multifrequency species classification of acoustic-trawl survey data using semi-supervised learning with class discovery. J Acoustical Soc Am 131(2):184–190
Xu P, Brock GN, Parrish RS (2009) Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. Comput Stat Data Anal 53(5):1674–1687
Zhang F (2006) The Schur complement and its applications. Springer, New York
Zhou Z, Wagner A, Mobahi H, Wright J, Ma Y (2009) Face recognition with contiguous occlusion using markov random fields. In: 2009 IEEE 12th international conference on computer vision, pp 1050–1057