Proceedings of the 5th International Conference on Statistics: Theory and Applications (ICSTA'23)
Brunel University, London, United Kingdom – August 03-05, 2023
Paper No. 148
DOI: 10.11159/icsta23.148

Variable Selection for a Contaminated Mixture of Normals Classification Model

Jorge Sánchez[1], Nema Dean[2], Tereza Neocleous[3]
[1]University of Glasgow
132 University Pl, Glasgow, G12 8TA
j.sanchez.1@research.gla.ac.uk; nema.dean@glasgow.ac.uk; tereza.neocleous@glasgow.ac.uk

Abstract **-** The problem of classifying observations into classes in the presence of contamination in high dimensions is addressed by combining variable selection and mixtures of contaminated normal distributions. Model-based (contaminated) discrimination is wrapped in a forward head-long variable search algorithm to identify and select separating variables. A simulation study carried out suggests that using only the selected variables set produces better classification performance than using only the true separating variables or all variables for balanced and unbalanced classes. This effect is more evident with data sets in high dimensions.

Keywords**:** Mixture of contaminated normals, variable selection, classification problems in high-dimensional space, supervised learning, accuracy.

1. Introduction

Mixture models have often been used in classification problems of all types [1]. These models represent the data with a finite mixture of distributions, each one corresponding to a class. The number of classes is assumed to be known, as well as the absence of contaminated samples. However, the assumption of only non-contaminated samples may be unrealistic in practice.

The presence of contaminated samples can have a negative impact on the accuracy of classification models since contamination in the training data affects the estimation of the parameters of the model [2][3]. McNicholas (2017) introduced a mixture of two normal distributions where an additional component with two parameters is used to model contamination for continuous data [4]. One of the extra parameters models the percentage of non-contaminated samples in the group and the other models an inflation variance factor related to contamination in each class. Unfortunately, under this modeling framework, it can be difficult if not impossible to deal with data sets in high-dimensional space.

Bouveyron (2013) states that model-based methods show disappointing behavior in high-dimensional spaces due to over-parametrization [5]. Moreover, there are several applications such as analytical spectroscopy, mass spectroscopy, or genomics where the number of available observations can be small compared with the number of variables, making parameter estimation even more difficult. Nevertheless, it is often possible to reduce the dimension of the original space due to the fact that the dimension of observed data is usually higher than the intrinsic dimension. In other words, a small number of variables contain the relevant information of the observations and the remainder contain irrelevant information.

Recent developments in model-based classification such as regularization-based techniques, parsimonious modeling, subspace classification methods, and classification methods based on variable selection enable efficient classification in high dimensional spaces. The advantage of the variable selection approach is that it can help with interpretability and it produces a simpler model. A simple model requires fewer parameters and because of this, it can make it possible to apply it in cases where it would not be possible using all the variables.

In this paper, we wrap a mixture of contaminated normal models in a forward head-long variable search algorithm [6] to enable classification in the presence of contamination in high dimensional spaces. This is based on combining the methods proposed by Punzo (2018) [7] and Dean (2006) [8] on mixtures of contaminated normal distributions and variable selection in model-based clustering of continuous variables.

In Section 2 we review some aspects of basic classification with and without contamination using a finite mixture of normal distributions. The model selection procedure is also described as well as the metric to assess models. Moreover, the forward head-long variable search algorithm is introduced and the steps to extend a mixture of contaminated normal distributions for classification problems in the presence of contamination to a high dimensional space. In Section 3 we describe our methodology to simulate different data sets taking into account different effects and evaluate our proposed approach. Finally, conclusions can be found in Section 4.

2. Methods

2.1. Classification

Classification is the process of allocating group membership labels to unlabelled observations. The presence of prior knowledge of group membership for observations (i.e. labelled data) can be used in classification problems. Moreover, in the finite mixture model framework `model-based discriminant analysis' can be used for supervised classification [1].

group

2.2. Finite Mixture Models

A finite mixture model is a probabilistic model which represents the presence of G groups within an overall population $\boldsymbol{X}$ of size n where p variables are observed for each subject. Finite mixture models have been broadly used in clustering and classification problems. So, each unlabelled element $\boldsymbol{x}$ of the population $\boldsymbol{X}$ is a random vector of dimension p coming from a parametric finite mixture distribution with density:

$$f(\boldsymbol{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\boldsymbol{x}|\boldsymbol{\theta_g}) \tag{1}$$

where $\boldsymbol{\vartheta} = (\pi_1, \dots, \pi_G, \boldsymbol{\Theta})$ is the vector of parameters for the mixture model, $\pi_g > 0$, is the proportion of the g^{th} component, $\boldsymbol{\Theta} = (\boldsymbol{\theta_1}, \dots, \boldsymbol{\theta_G})$ the vector of parameters of the component densities, $\sum_{g=1}^{G} \pi_g = 1$, and $f_x(\boldsymbol{x}, \boldsymbol{\theta_g})$ the density function of the g^{th} component.

2.3. Mixture of multivariate Gaussian distributions

The most popular finite mixture model used in classification problems is the Gaussian mixture model with g^{th} component density:

$$f_g(x|\boldsymbol{\theta_g}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma_g}|}} \mathrm{e}^{\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_g})^T \boldsymbol{\Sigma_g^{-1}} (\boldsymbol{x}-\boldsymbol{\mu_g})\right)}, \tag{2}$$

where $\boldsymbol{x} \epsilon \mathbb{R}^p$ has a multivariate Gaussian distribution with parameters $\boldsymbol{\theta_g} = (\boldsymbol{\mu_g}, \boldsymbol{\Sigma_g})$ with mean vector $\boldsymbol{\mu_g}$ and covariance matrix $\boldsymbol{\Sigma_g}$, we can observe that a general Gaussian mixture model contains a total of $(G-1) + Gp + \frac{Gp(p-1)}{2}$ parameters and as p increases the number of parameters to estimate in the covariance matrices to estimate also rapidly increases, limiting the application of this probabilistic model.

2.4 Model-based discriminant analysis

We assume a data set composed of n p-dimensional vectors $\boldsymbol{x_1}, \dots, \boldsymbol{x_n}$ is observed and we know the group labels for these observations. Let $\boldsymbol{l_i} = \{l_{i1}, \dots, l_{iG}\}$ for $i = 1, \dots, n$ be the associated label, where $l_{ig} = 1$ if the observation belongs to class g and 0 otherwise. The model-based discriminant analysis likelihood can be expressed as follows:

$$l(\boldsymbol{\vartheta}) = \prod_{i=1}^{n} \prod_{g=1}^{G} \left[\pi_g N(\boldsymbol{x_i}, \boldsymbol{\mu_g}, \boldsymbol{\Sigma_g})\right]^{l_{ig}}, \tag{3}$$

The proportion π_g is seen as the prior probability that an observation belongs to the class g.

2.5 Mixture of contaminated Gaussian distributions

In many real data situations, the presence of contamination in a data set is very likely. In this case, we can model each class by a mixture of two normal distributions: one for non-contaminated $N(\boldsymbol{x_i}, \boldsymbol{\mu_g}, \boldsymbol{\Sigma_g})$ and for contaminated $N(\boldsymbol{x_i}, \boldsymbol{\mu_g}, \eta_g \boldsymbol{\Sigma_g})$, where the proportion of non-contaminated observations is $\alpha_g \epsilon\ (0,1)$ and the inflation of variance factor, to allow higher variability around the mean for contaminated observations, is given by $\eta_g > 1$ [3]. We define $\{\boldsymbol{v_i}\}_{i=1}^{n}$, $\boldsymbol{v_i} = (v_{i1}, \dots v_{ig})$ as unobserved labels indicating non-contaminated data versus contaminated data for each group, where $v_{ig} = 1$ if observation i of the class g is non-contaminated and $v_{ig} = 0$ if observation i in class g is contaminated. The complete data-likelihood and log-likelihood are given by:

$$l(\boldsymbol{\vartheta}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \prod_{i=1}^{n} \prod_{g=1}^{G} \left[\Pi_g \left[\alpha_g N(\boldsymbol{x_i}, \boldsymbol{\mu_g}, \boldsymbol{\Sigma_g}) \right]^{v_{ig}} \left[(1 - \alpha_g) N(\boldsymbol{x_i}, \boldsymbol{\mu_g}, \eta_g \boldsymbol{\Sigma_g}) \right]^{(1 - v_{ig})} \right]^{l_{ig}}, \tag{4}$$

$$log(l(\boldsymbol{\vartheta}, \boldsymbol{\alpha}, \boldsymbol{\eta})) = \sum_{g=1}^{G} \sum_{i=1}^{n} l_{ig} \left[log \pi_g + v_{ig} \log \left(\alpha_g N(\boldsymbol{x_i}, \boldsymbol{\mu_g}, \boldsymbol{\Sigma_g}) \right) + (1 - v_{ig}) \log((1 - \alpha_g) N(\boldsymbol{x_i}, \boldsymbol{\mu_g}, \boldsymbol{\Sigma_g})) \right] . \tag{5}$$

As mentioned in Section 2.3 the number of parameters to be estimated increases rapidly with the increase in the number of variables and as a result, the mixture of contaminated Gaussian distributions cannot deal with data in high-dimensional space. Nevertheless, this framework can be extended to manage data sets in high-dimensional spaces by wrapping a mixture of contaminated Gaussian models in a headlong search algorithm to select the relevant variables.

2.6 Model selection

To evaluate and select models, a data set with *n* observations is randomly split into subsets which are named training and test sets with *m* and *n* – *m* observations respectively. We know the group information for all *n* observations. However, we pretend that we only know the group information for *m* observations in the training set and not for the *n* – *m* observations in the test set. The training set is used to estimate the parameters of the model(s), while the test set is used to evaluate the model(s) with observations that were not part of the training to assess how the model(s) would perform with new observations. Let $\mathbf{z_i}$ be a realization of $\mathbf{Z_i}$ for $i = m + 1, \dots, n$ which is a random variable that follows a multinomial distribution with one draw on G categories with probabilities $\pi_1, \dots, \pi_g$. Moreover, $\mathbf{Z_i} = \{Z_{i1}, \dots, Z_{iG}\}$ are assumed to be independent and identically distributed. The posterior probability can be expressed as follows:

$$\Pr[Z_{ig} = 1 | \boldsymbol{X}_i] = \frac{\left[\pi_g \left[\alpha_g N(\boldsymbol{x_i}, \boldsymbol{\mu_g}, \boldsymbol{\Sigma_g}) \right]^{v_{ig}} + \left[(1 - \alpha_g) N(\boldsymbol{x_i}, \boldsymbol{\mu_g}, \eta_g \boldsymbol{\Sigma_g}) \right]^{(1 - v_{ig})} \right]}{\sum_{h=1}^{G} [\pi_h [\alpha_h N(\boldsymbol{x_i}, \boldsymbol{\mu_h}, \boldsymbol{\Sigma_h})]^{v_{ih}} + [(1 - \alpha_h) N(\boldsymbol{x_i}, \boldsymbol{\mu_h}, \eta_h \boldsymbol{\Sigma_h})]^{(1 - v_{ih})}]}, i = m + 1, \dots, n \tag{6}$$

Once the parameters are estimated, then the estimated group membership is given by

$$\hat{z}_{ig} = \frac{\left[\hat{\pi}_g \left[\hat{\alpha}_g N(\boldsymbol{x_i}, \hat{\boldsymbol{\mu}}_{\boldsymbol{g}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{g}}) \right]^{\hat{v}_{ig}} + \left[(1 - \hat{\alpha}_g) N(\boldsymbol{x_i}, \hat{\boldsymbol{\mu}}_{\boldsymbol{g}}, \hat{\eta}_g \hat{\boldsymbol{\Sigma}}_{\boldsymbol{g}}) \right]^{(1 - \hat{v}_{ig})} \right]}{\sum_{h=1}^{G} \left[\hat{\pi}_h \left[\hat{\alpha}_h N(\boldsymbol{x_i}, \hat{\boldsymbol{\mu}}_{\boldsymbol{h}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{h}}) \right]^{\hat{v}_{ih}} + \left[(1 - \hat{\alpha}_h) N(\boldsymbol{x_i}, \hat{\boldsymbol{\mu}}_{\boldsymbol{h}}, \hat{\eta}_h \hat{\boldsymbol{\Sigma}}_{\boldsymbol{h}}) \right]^{(1 - \hat{v}_{ih})} \right]}, i = m + 1, \dots, n \tag{7}$$

It is clear that each observation has a posterior probability of belonging to each class. For example, if we have two classes then the observation will be a vector composed of two elements which are the posterior probabilities of belonging to each class. Let us assume we have two classes and the observation *i* has following posterior probabilities $\boldsymbol{z}_i = (0.2, 0.8)$. There are cases where we need to allocate an observation to one class, so we allocate it to the class with the maximum posterior probability (MAP). In the above case $MAP(0.2, 0.5) = (0,1)$, i.e. the i^{th} observation is assigned to the 2^{nd} class.

In general,

$$MAP(\hat{z}_{ig}) = \begin{cases} 1, & if\ g = \arg max_h\ \hat{z}_{ih} \\ 0, & otherwise \end{cases} \quad (8)$$

When there are different candidate models, a model selection criterion is needed. There are many different metrics proposed to choose a model among different candidate classification models. Very common metrics used to choose models are accuracy, sensitivity, specificity, F1-score, area under the curve (AUC), and others. These metrics are usually calculated over the test set. We define true positive TP_g as the number of observations predicted in class g that actually belong to class g. Similarly, we define true negative TN_g as the number of observations predicted not to be in class g that actually do not belong to class g and false positive FP_g as the number of observations predicted in class g that actually belong to a different class (false negatives FN_g analogously). In a two-class case (positive and negative classes), a cross classification table is shown in Table 1.

Table 1: Cross classification table for two class problem

Predicted	Actually Positive	Actually negative	Total
Positive	TP	FP	TP + FP
Negative	FN	TN	FN + TN
Total	TP + FN	TP + TN	Total (*=n-m)*

Accuracy for a classification method is calculated by dividing the number of correctly predicted labels by the total number of observations in the test set.

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} \quad (9)$$

2.7 Wrapping a mixture of contaminated normal distribution in a forward head-long search algorithm

We propose using the headlong forward variable search algorithm (Nielsen et al, 1998) for the contaminated discriminant analysis model. This search approach (with a different classification model) has been used previously in spectroscopic data sets with success [6]. The steps of the headlong forward variable search algorithm as follows:

1. Order the variables in terms of initial classification power calculating their univariable F-statistics based on the g groups for the training data in increasing order for the variables. (They remain in this order for all following steps unless removed).
2. To choose the first variable to be included in the proposed model, do a greedy search of one-dimensional proposed models of the variables, calculating accuracy (9) for each one-dimensional model and choosing the variable which gives us the highest accuracy on the test set to obtain the proposed model that will be our new current model.
3. The general inclusion step is decided using the sets of selected variables and non-selected variables. Check one variable at a time from the non-selected variables set for inclusion in the proposed model and if the accuracy on the test set improves over the current model, update the current model by adding this proposed variable to the set of selected variables and stop.
4. Repeat the inclusion step. Stop the search algorithm if at any stage we get to the end of the list of non-selected variables without improvement or if all variables are selected.

3 Simulation study

We propose a framework for simulating different scenarios and assessing proposed models. In this framework, we varied the distance between group means, number of groups, class proportions, number of variables and type of pairwise correlation. In all cases there were 2 true separating variables with the remainder being non-separating. We considered the following settings (see Table 2).

Table 2: Different factor settings in simulation framework.

Factors	Description	Levels
V	Set of variables to train the model	True variables, all variables, variables selected by headlong search
F1	The distance between mean classes	Very overlapping (VO) (difference = 1.5σ) Medium distance (MD) (difference = 3σ) Very distant (VD) (difference = 6σ)
F2	Number of groups	2, 3
F3	Class proportion	Balanced (0.5/0.5) Unbalanced (0.9/0.1)
F4	Number of variables	4,100
F5	Correlation structure	Strong correlation between separating variables (SCBSV) ($\rho = 0.8$) Strong correlation between separating and no separating variables (SCBSNSV) ($\rho = 0.8$) Strong correlation between no separating variables (SCBNSV) ($\rho = 0.8$) Independence (IND) ($\rho = 0$)

The number of total simulations per setting is 100 and the number of simulated settings is 92 which means an overall count of 9,200 simulated data sets.

To have an idea of what the potentially influential factors are in terms of accuracy, we looked at box plots for each factor across all other factors in 4 dimensions (see Figure 1). In this example, half of the variables (X_2 and X_4) contain group information while the other half are noisy variables (X_1 and X_3) and we can observe that using the model that uses the selected variables produces better performance in comparison with including the true variables or all the variables. It seems that the accuracy is affected negatively by the number of classes, a strong correlation between separating variables in the covariance structure, and a very overlapping distance between group means. It is expected that having a very overlapping distance between groups would negatively impact the performance of classification models since this represents a more difficult classification problem. Consequently, if our proposed approach can deal with this scenario in 4 dimensions, where half of the variables are noisy and in a higher dimension (100), where 98% of the variables are noisy, then we might assume that it performs better in data sets where class separation is clearer. Hence, it is plausible to explore more closely settings with very overlapping class mean distances.

3.1 Classification in the presence of contamination for two group balanced data sets with very overlapping group means

We look at data sets with very overlapping group means in low and higher dimensions and with different covariance structures to observe the behaviour of our proposed approach. The groups are balanced and the performance of models is given by their accuracy in the test set. We expect to see a higher mean accuracy using the selected variables for the test set

in comparison with using true variables and all variables, especially in higher dimensions.

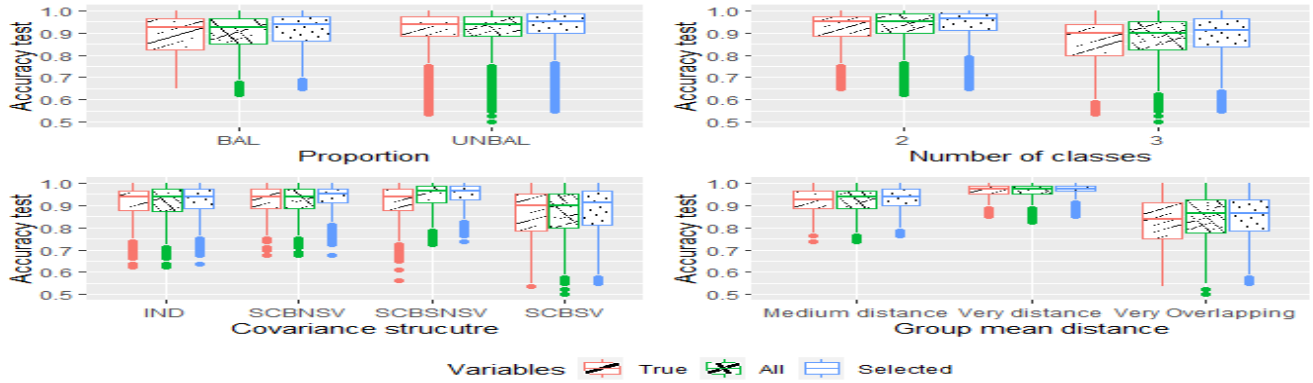

Fig. 1: Accuracy by sets of variables for settings with variation in the following factors F1,F2,F3, and F5 (see Table 2) in 4 dimensions

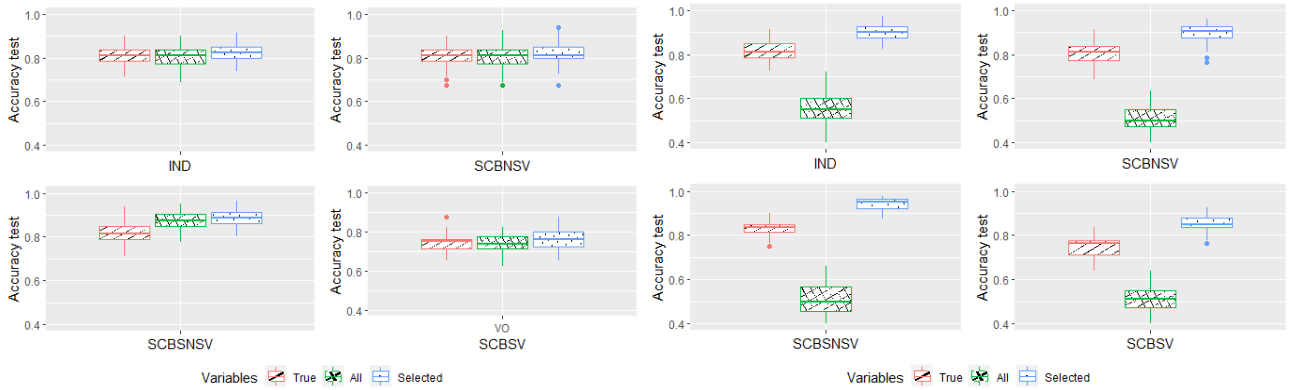

Fig. 2: Accuracy by sets of variables for two balanced groups mapped in 4 dimensions (left) and 100 dimensions (right) for different covariance structures

3.1.1 Balanced data sets in 4 dimensions

We can observe the accuracy of two balanced groups in 4 dimensions for different covariance structures (see Figure 2). We notice that adding all variables has a negative impact on the accuracy. However, the use of the selected variables improve the mean accuracy by an average of 2% and 5% in comparison with using all variables and the true variables.

3.1.2 Balanced data sets in 100 dimensions

If the data set is in higher dimensions, for example in 100 variables the increase in the average accuracy obtained using the selected variables is larger in comparison with the other two sets of variables for all covariance structures. We notice that adding all variables harms the accuracy. Nevertheless, the use of the selected variables improves the mean accuracy of the model by an average of 32% and 10% in comparison with using all the variables and the true variables (see Figure 2).

3.2 Classification in the presence of contamination in unbalanced datasets

We repeat the same analysis but considering 2 unbalanced groups in 4 and 100 dimensions to evaluate how the average accuracy using different sets of variables might be impacted.

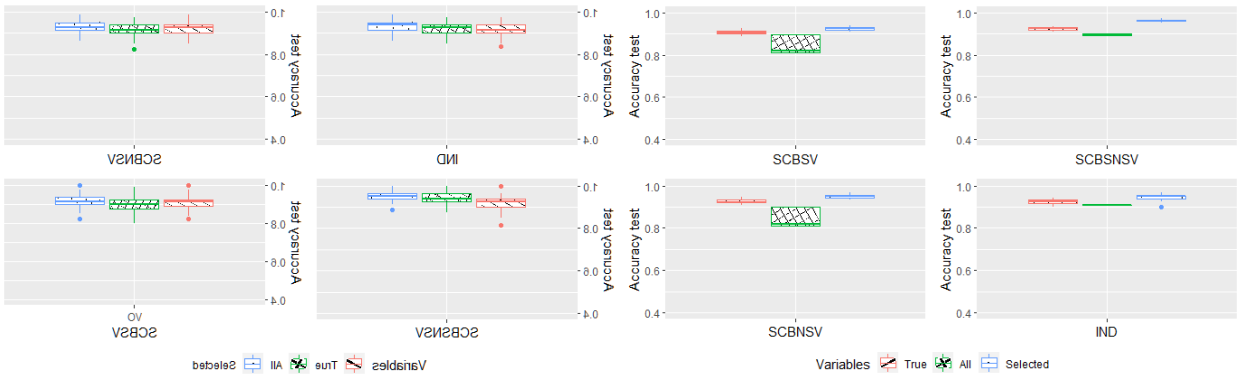

Fig. 3: Accuracy by sets of variables for two unbalanced groups mapped in 4 dimensions (left) and 100 dimensions (right) for different covariance structures.

3.2.1 Unbalanced data sets in 4 dimensions

It seems that in these simulated data sets the average accuracy improves using the selected variables for all covariance structures for data sets in 4 dimensions (see Figure 3). Moreover, the use of the selected variables produces classification models with slightly better performance in terms of mean accuracy by an average of 2.1% and 2.3% in comparison with using all the variables and the true variables. Nevertheless, there is not a large difference between using all variables and the true variables. Using all variables harms the mean accuracy by an average of 0.3% in comparison with using the baseline (the true variables).

3.2.2 Unbalanced data sets in 100 dimensions

When we revisit the previous scenario in a higher dimension, we observe in Figure 3 that when the number of noisy variables increases, it is clearer that using the selected variables produces classification models with better performance in terms of mean accuracy by an average of 8% and 3% in comparison with using all the variables and the true variables.

3.3 Modeling accuracy by factors

Suppose we are measuring the accuracy y of a classification model applied to simulated data coming from the settings described (see Table 2). For each setting 100 data sets are simulated and for each simulation, the accuracy of the models for three different sets of variables (true separating variables, all variables and variables selected by headlong search) is recorded.

3.3.1 Modeling accuracy with a linear mixed effects model

We start by looking at accuracy as a function of the factors describing the data sets (Table 2). The model is expressed as:

$$y_{ij} = \beta_0 + \alpha_1 V_{ij} + \beta_1 F_{1i} + \beta_2 F_{2i} + \beta_3 F_{3i} + \beta_5 F_{5i} + b_i + \varepsilon_{ij} \quad (10)$$

where

y_{ij} accuracy measurement for the i^{th} simulation and j^{th} set of variables, i = 1, …, 19,200 and j = 1,…,3

β_0 intercept

V_{ij} j^{th} set of variables used in the i^{th} $simulation$

$F1_i$ distance between mean classes for the i^{th} simulation

$F2_i$ number of classes for the i^{th} simulation

$F3_i$ class proportion for the i^{th} simulation

$F5_i$ covariance structure for the i^{th} simulation

b_i random effects for simulation where $b_i \sim N(0, \sigma_b^2)$

ε_{ij} error for the where $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$

We can observe that for data sets with 4 variables all main effects are significant and that group mean distance and number of classes factors have the largest effect on accuracy (Table 3). In addition, increasing the number of classes to having a covariance structure with strong covariance between separating variables SCBSV, and a group mean distance overlapping (VO) contribute negatively to accuracy (Table 3). We checked the residual plots for violations of the assumption and were satisficatory.

Table 3: Coefficient estimates of the model for accuracy.

Sources	Estimate	p-value
Intercept	0.91	<.0001
Variables – Selected	0.01	<.0001
Variables – True	-0.00	<.0001
Proportion – Unbalanced	0.06	<.0001
Number of classes – 3	-0.09	<.0001
Covariance Structure – SCBNSV	0.02	<.0001
Covariance Structure – SCBSNSV	0.02	<.0001
Covariance Structure – SCBSV	-0.04	<.0001
Group Mean Distance – VD	0.04	<.0001
Group Mean Distance – VO	-0.09	<.0001

4. Conclusion

The result of simulations suggests that when we add more noisy variables, the performance of a classification model with balanced or unbalanced data using the selected variables will be better than using the true separating variables or all variables. Using all the variables is problematic because this limits the cases we can apply the model due to the number of observations needed to estimate the large number of parameters resulting from high dimensions. In addition, we will never know the true variables that separate groups in practice. The proposed approach seems to be useful in different settings, especially in higher dimensions where there is the presence of contaminated samples and a huge number of noisy variables.

Acknowledgements

The first author was supported by a scholarship awarded by The Secretariat of Higher Education, Science, and Technology and Innovation of Ecuador (Agreement No. SENESCYT-SDFC-DSEFC-2018-0215-O).

References

[1] G. J. McLachlan, *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons, 2005.

[2] A. Cappozzo, F. Greselin and T. Murphy, "*A robust approach to model-based classification based on trimmings and constraints: Semi-supervised learning in presence of outliers and label noise."* Advances in Data Analysis and Classification, 2020, vol. 14(2), pp. 327-354.

[3] A. Punzo and P. McNicholas, *"Parsimonious mixtures of multivariate contaminated normal distributions"*, Biometrical Journal, vol. 58, no. 6, pp. 1506-1537.

[4] P. McNicholas, *"Mixture model-based classification"*, CRC, 2016.

[5] C. Bouveyron, "Probabilistic model-based discriminant analysis and clustering methods in chemometrics" Journal of Chemometrics., vol. 83, no. 4, pp. 433-446, 2013.

[6] T. Murphy, N. Dean, A. Raftery. *"Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications."* Annuals of Applied Statistics, vol. 4, no.1, p. 396, 2010.

[7] A Punzo, A. Mazza, P. McNicholas, *"ContaminatedMixt: An R package for fitting parsimonious mixture of multivariate contaminated normal distributions."* arXiv preprint arXiv:1606.03766, 2016.

[8] N. Dean, T. Murphy, and G. Downey, *"Using unlabelled data to update classification rules with applications in food authenticity studies."* Journal of the Royal Statistical Society Series C, vol. 55, no.1, pp.1-14,2006.

[9] N. Neykov, P. Filzmoser, *"Robust fitting of mixtures using the trimmed likelihood estimator"* Computational Statistics and Data Analysis, vol. 52, no.1, pp.299-308,2007.