Imprecise Gaussian Discriminant Classification

Yonatan Carlos Carranza-Alarcon, Sébastien Destercke

Imprecise Gaussian Discriminant Classification*

Yonatan-Carlos Carranza-Alarcon[a], Sébastien Destercke[a]

[a] *UMR CNRS 7253 Heudiasyc, Sorbonne universités, Université de technologie de Compiègne CS 60319 - 60203 Compiègne cedex, France*

Abstract

Gaussian discriminant analysis is a popular classification model, that in the precise case can produce unreliable predictions in case of high uncertainty (scarce or noisy data set). While imprecise probability theory offer a nice theoretical framework to solve such issues, it has not been yet applied to Gaussian discriminant analysis. This work remedies this, by proposing a new Gaussian discriminant analysis based on robust Bayesian analysis and near-ignorance priors. The model delivers *cautious* predictions, in form of set-valued class, in case of limited or imperfect available information. Experiments show that including an imprecise component in the Gaussian discriminant analysis produce reasonably cautious predictions, and that set-valued predictions correspond to instances for which the precise model performs poorly.

Keywords: Discriminant Analysis, Robust Bayesian, Classification, Near-Ignorance

1. Introduction

In machine learning, the classification task consists in seeking to identify to which label (among a finite set $\mathscr{K}$ of such labels) a new unlabelled instance $\boldsymbol{x} \in \mathscr{X}^p$ belongs. The reliability of this *precise* prediction (or *single* decision) may depend heavily on prior beliefs (e.g. assumptions made by data analysts, such as asymptotically unbiased estimators) and the nature of training data sets (e.g. in small amounts [1, 2] and/or with high degree of *uncertainty*[1]), both will be referred as *imperfect information*[2]. A well-known *precise* generative classifier model used to perform the classification task is the Gaussian discriminant analysis (GDA) [6, §4.3]

A classifier model is called *precise* when it performs pointwise predictions (or precise estimations) in the form of single class labels, even in extreme cases, regardless of the available information we have a about an instance. In these cases, it may be useful to provide set-valued, but more reliable predictions, especially for sensitive applications (e.g. medical diagnosis, control systems, etc.) where we cannot afford to make mistakes (see illustration in Figures 3(a) and 3(b)).

Imprecise probabilities (IP)[7] can mitigate the impact of imperfect information, taking into account the lack of evidence by replacing *precise* estimates (or a single probability distribution) with *imprecise* estimates (or a set of probability distributions, most often in the form of convex set) in order to make *cautious* set-valued decisions. In this paper, we adopt such an approach to propose a novel method of (cautious[3]) *imprecise classification*.

Cautious classification is a relatively new trend in machine learning which do not aim to do "better" than their precise counterparts, nor to implement a rejection option (i.e., not classifying at all) in case

*This paper is part of the published paper in Logic Fussy and its Applications (LFA-2018).

Email addresses: yonatan-carlos.carranza-alarcon@hds.utc.fr (Yonatan-Carlos Carranza-Alarcon), sebastien.destercke@hds.utc.fr (Sébastien Destercke)

[1] *Uncertainty* can be due to lack of knowledge or to the natural variability in the observed data [3, ch.2] (or a.k.a. *epistemic and aleatoric uncertainty* [4]), and it can lead us to biased estimations and high variance models [5].

[2] Imperfect information is here used as a synonym for limited information or/and lack of knowledge or prior beliefs.

[3] Cautious and imprecise are here used interchangeably.

of ambiguity [8], but to highlight those hard cases for which information is insufficient to isolate a *single* reliable precise prediction, and to propose a subset of possible predictions. We can find in the literature three "main" ways to access cautious classifier models: (1) using a classical precise classifier but deriving a set-valued predictions from them [9] (e.g. partial reject [10], conformal prediction [11]) (2) making data imperfect (coarse or impartial observations) and then building a corresponding imperfect robust model, and finally (3) a cautious classifier under IP from which set-valued predictions follow naturally (such as robust frequentist inference [12, 13] or Bayesian inference [14, 15, 16]). We retain here the latter as it considers imprecision as parts of its basic axioms, rather than the other approaches where imprecision is added ex-post.

Bayesian methods incorporate some prior beliefs in the form of probability distribution defined on unknown parameters of the model. Such beliefs typically comes from expert opinions of persons that are knowledgeable in the context of the problem. However, it is also well-known that the elicitation of prior beliefs can be absent or hard to obtain during the study of a problem, especially in when learning classifier. A classical way out of this problem is to use so-called non-informative prior, that allow one to obtain a posterior not including any prior knowledge [2]. Yet, the use of such prior is not without problem within the Bayesian theory, as they are not coherent in the sense of De Finetti. In addition, it has been argued and shown that using truly vacuous prior information while remaining coherent usually lead to vacuous posterior predictions [14, 17, §7.4, §5.6.2] (i.e. our model would not be able to learning from data). Moreover, it may seem strange that an absence of prior should lead to a fully precise, completely informed posterior. Walley have therefore proposed to use a set of non-informative prior distributions, called *near-ignorance prior* [14, §4.6.9], to solve this issue. These near-ignorance prior must respect certain properties [18, §2] so as not to obtain vacuous predictions. Hence, one of our motivations in this paper is to not to use a single prior distribution, but a set of prior distributions (or credal set [19]) to reflect our lack of knowledge and obtain cautious predictions.

Let $\mathscr{X} \times \mathscr{K}$ be the space of observations and possible labels, with$X \in \mathscr{X} = \mathbb{R}^p$ a random vector and $Y \in \mathscr{K} = \{m_1, ..., m_K\}$ the set of categories. The main goal of GDA is to estimate the theoretical conditional probability distribution (c.p.d) $\mathbb{P}_{Y=m_k|X}$ of the class $Y = m_k$ given the observation X via Bayes' theorem as follows

$$\mathbb{P}_{Y=m_k|X} = \frac{\mathbb{P}_{X|Y=m_k}\mathbb{P}_{Y=m_k}}{\sum_{m_l \in \mathscr{K}} \mathbb{P}_{X|Y=m_l}\mathbb{P}_{Y=m_l}}. \tag{1}$$

Thus, quantifying $\mathbb{P}_{Y=m_k|X}$ is equivalent to quantify $\mathbb{P}_{X|Y=m_k}$ and the marginal distribution $\mathbb{P}_Y$. In *precise* probabilistic approaches, this is typically done by using maximum likelihood estimation (MLE) and by making some parametric assumptions about the probability density $\mathbb{P}_{Y=m_k|X}$ (i.e. Gaussian probability distribution (g.p.d)) in order to find a plausible estimate (see Section 3.1). However, such precise estimates usually have trouble differentiating different kinds of uncertainties [4], such as uncertainty due to ambiguity (mixed classes in some areas of the input space) and uncertainty due to lack of knowledge or information (limited training data set inducing biases in estimates [5]). In both cases, it may be useful to provide set-valued, but more reliable predictions, especially for sensitive applications where we cannot afford to make mistakes (see illustrations in Figures 3(a) and 3(b)).

Section 2 recalls the basics of the *precise* classification setting, adopting the viewpoints of statistical decision theory [6, §2] and expected utility [20, §2.2]. It also introduces the corresponding extensions of these tools to the IP setting, in particular the *maximality* criterion [21] that extend the *precise* utility-based decision-making to the IP context.

In Section 3, we describe the estimation of the conditional distribution in the case of the precise GDA, using a frequentist inference approach. We then extend this *precise* parametric estimation to *imprecise* estimation in a robust Bayesian inference context, using the IP near-ignorance model proposed by Benavoli *et al.* [18] to do so and obtaining estimates in the form of a convex set . Coupling this imprecise estimation with the *maximality* criterion, we present our Imprecise GDA (IGDA) model and its different variants in Section 4.

In Section 5, we perform a set of experiments on different datasets using our imprecise model and compare it to its precise counterparts. We show that the cautious predictions are useful, in the sense that they concern instances for which the precise classifier often makes mistakes, often include the true class within

the predicted set or these same instances, and are not overly imprecise. Furthermore, we briefly discuss (focusing on computational issues) in Section 6 the extension of our method to other settings, namely to the case where the class proportions $\mathbb{P}_{Y=m_k}$ are also imprecisely estimated, and where the criteria to minimise is not the raw number of errors (corresponding to a 0/1 loss function) but a generic loss function. In this paper, we will use mathematical notations of table 1.

Symbols	Description
Data related notations	
$\mathscr{X}^p \subseteq \mathbb{R}^p$	Input space of dimension p.
$\mathcal{K} = \{ m_1, ..., m_K \}$	Output space of size K.
$(\cdot)^T$	Transpose Operator
$\|\cdot\|$	Euclidian norm
$\boldsymbol{X}$	Matrix $n \times p$ of all instances of dataset.
$\boldsymbol{y} = (y^1, \ldots, y^n)^T$	Vector $n \times 1$ of all label of dataset.
$\boldsymbol{x} = (x^1, \ldots, x^p)^T$	New unlabeled instance to predict.
$\hat{y}$	Precise output prediction.
$\hat{Y}$	Set-valued output predictions.
N	Number of training instances.
Decision theory	
Θ	Parameter space
$\mathscr{P} = \{\mathbb{P}_\theta,\ \theta \in \Theta\}$	Set of probability distributions
$\mathbb{P}_{Y\|X}, \mathbb{P}_{X\|Y}, \mathbb{P}_Y$	(Conditional) Probability distributions
$P(Y = m_k \| X = \boldsymbol{x}) \sim \mathbb{P}_{Y\|\boldsymbol{x}}$	Conditional probability of m_k given $\boldsymbol{x}$
$\Phi = \{\varphi \ \| \varphi : \mathscr{X} \times \Theta \to \mathcal{K}\}$	Learning space models
$\mathscr{L}_{0/1}(y, \varphi(\boldsymbol{x})) = \mathbb{1}_{y \neq \varphi(\boldsymbol{x})}$	*zero–one* loss function
Discriminant Analysis	
n_k	Number of observations of category m_k
$(\boldsymbol{x}_{i,k}, y_{i,k})_{i=1}^{n_k} = \{(\boldsymbol{x}_{1,k}, y_{1,k}), \ldots, (\boldsymbol{x}_{n_k,k}, y_{n_k,k})\}$	Observations of category m_k.
$\overline{\boldsymbol{x}}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \boldsymbol{x}_{i,k}$	Empirical mean of category m_k
$\widehat{\sigma}^j_{m_k} = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (x^j_{i,k} - \overline{x}^j_k)^2, \forall j \in \{1, \ldots, p\}$	Empirical variance of category m_k
$\widehat{S}_{m_k} = \frac{1}{N - n_k} \sum_{i=1}^{n_k} (\boldsymbol{x}_{i,k} - \overline{\boldsymbol{x}}_k)(\boldsymbol{x}_{i,k} - \overline{\boldsymbol{x}}_k)^T$	Empirical covariance matrice of category m_k
$\widehat{S} = \frac{1}{(N-K)} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\boldsymbol{x}_{i,k} - \overline{\boldsymbol{x}}_k)(\boldsymbol{x}_{i,k} - \overline{\boldsymbol{x}}_k)^T$	Empirical total covariance matrix
$\widehat{\pi}_y = \{\widehat{\pi}_{y=m_k} \mid \widehat{\pi}_{y=m_k} = n_k/N, \sum_{m_k \in \mathcal{K}} \widehat{\pi}_{y=m_k} = 1\}$	Empirical marginal distribution $\mathbb{P}_Y$.

Table 1: Mathematical notations used in this paper

2. Preliminaries and basic reminders

In this section, we remind some notions of *classical* statistical learning and decision-making used to build a *precise* classification model, as well as basic notions needed to also deal with sets of probabilities.

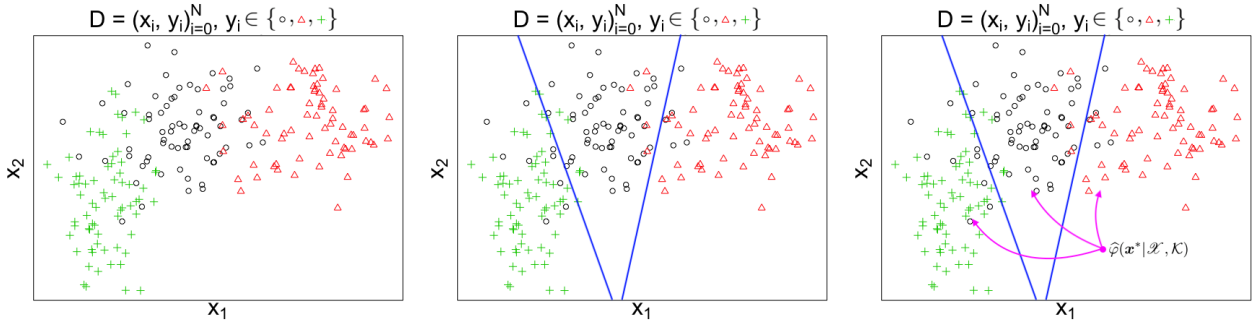
2.1. Classification setting

Let $\mathscr{D} = \{(\boldsymbol{x}_i, y_i) | i = 1, \ldots, N\}$ be a training data set issued from $\mathscr{X}^p \times \mathcal{K}$, such that $\boldsymbol{x}_i \in \mathscr{X}^p$ are regressors or features (input space) and $y \in \mathcal{K}$ is the response variable or class (output space). We denote n_k the number of observations that belong to the label m_k, and so $N = \sum_{k=1}^{K} n_k$.

The goal of classification is to build a predictive model $\varphi : \mathscr{X} \to \mathcal{K}$ that predicts a label $m_k \in \mathcal{K}$ given a new unlabelled instance $(\boldsymbol{x}, \cdot) \notin \mathscr{D}$. A well-known approach for that is called *inductive* learning which

involves two steps; (1) the learning phase that estimates or induce[4] a model from observed data, and (2) the decision inferred from the learned model.

Learning consists in determining the optimal model $\widehat{\varphi} \in \Phi$ among the chosen set Φ of models, using to do so a set of training data $(\boldsymbol{x}_i, y_i)_{i=1}^N$ generated from an unknown joint probability distribution $\mathbb{P}_{X,Y}$ (see figure 1 for an illustration). After getting an *"optimal"* model $\widehat{\varphi}$, we must decide what is the label of a new unlabelled instance $(\boldsymbol{x}, \cdot)$. This latter decision step can be handled by tools issued from decision theory.



(a) Getting Training Data (b) Learning model ϕ (c) Prediction unlabeled instance $\boldsymbol{x}$

Figure 1: *Learning model steps.* Figure (a) show the initial training data, from which are induced the boundaries defining the decision function (b), then used to perform the predictions (c).

In machine learning, induction is often seen as the task of determining a decision function that will minimise the risk of getting misclassifications. The cost or risk of a misclassification is generally quantified through a function $\mathscr{L} : \mathcal{K} \times \mathcal{K} \to \mathbb{R}$ penalising every bad decision, known as *loss* function, with $\mathscr{L}(y, \hat{y})$ the loss incurred when predicting $\hat{y}$ if y is the observed, true value. The optimal model is most often defined as the one minimizing the expected loss.

Definition 1 (Risk minimizing [6, §2.4])**.** *Given a general loss function $\mathscr{L}(\cdot, \cdot)$, the optimal model is defined as the one minimizing the average loss of getting missclassification.*

$$\hat{\varphi} := \operatorname{argmin}_{\varphi(X) \in \Phi} \mathbb{E}_{\mathscr{X} \times \mathcal{K}} \left[\mathscr{L}(Y, \varphi(X)) \right] \tag{2}$$

If loss function is defined instance-wise, then, Equation (2) can also be expressed as the minimization of conditional expectation [6, eq. 2.21]:

$$\hat{\varphi} := \operatorname*{arg\,min}_{y \in \mathscr{K}} \mathbb{E}_{Y|X} \left[\mathcal{L}(y, \varphi(X)) \right] \tag{3}$$

Classical accuracy corresponds to a *zero-one* loss function, where all missclassification are penalised identically, i.e. $\mathscr{L}_{0/1}(y, \hat{y})$ is equal to 1 if y and $\hat{y}$ are different and 0 otherwise. Therefore, given $\mathscr{L}_{0/1}$, we can reformulate the risk minimization as the well-known *Bayes classifier*, which would choose the learning model maximizing the conditional probability (a.k.a. maximum a posterior (MAP) probability) given a new unlabeled instance $\boldsymbol{x}$:

$$\widehat{\varphi}(\boldsymbol{x}) = \operatorname{argmax}_{m_k \in \mathscr{K}} P(Y = m_k | X = \boldsymbol{x}) \tag{4}$$

Hence, in a *precise* probabilistic approach, the main task is to estimate the conditional distribution $\mathbb{P}_{Y|X}$, from which can be obtained the optimal decision. An alternative way of looking at this decision-making problem is to pose it as a problem of inferring preferences between the labels, as follows:

Definition 2 (Precise ordering [20, pp. 47])**.** *Given a general loss function $\mathcal{L}(\cdot, \cdot)$ and a conditional probability distribution $\mathbb{P}_{Y|\boldsymbol{x}}$, m_a is preferred to m_b, denoted by $m_a \succ m_b$, if and only if:*

$$\mathbb{E}_{\mathbb{P}_{Y|\boldsymbol{x}}} \left[\mathcal{L}(\cdot, m_a) | \boldsymbol{x} \right] < \mathbb{E}_{\mathbb{P}_{Y|\boldsymbol{x}}} \left[\mathcal{L}(\cdot, m_b) | \boldsymbol{x} \right] \tag{5}$$

[4] In the sense that it goes from singular observations to a generic model.

Definition 2 tells us that exchanging m_b for m_a would incur a positive expected loss, due to the fact that expectation loss of m_b is greater than m_a, therefore m_a should be preferred to m_b for a given new unlabelled instance $\boldsymbol{x}$. In the particular case where we use the loss function $\mathscr{L}_{0/1}$, it is easy to prove that:

$$m_a \succ m_b \iff P(Y = m_a|X = \boldsymbol{x}) > P(Y = m_b|X = \boldsymbol{x}) \tag{6}$$

where $P(Y = m_a|X = \boldsymbol{x})$ is the unknown conditional probability of label m_a and a new unlabeled instance $\boldsymbol{x}$. Therefore, given a set of labels $\mathcal{K}$, we can then establish a complete preorder making pairwise comparisons (see figure 2) as follows:

$$m_{i_K} \succ m_{i_{K-1}} \succ \cdots \succ m_{i_1} \iff P(Y = m_{i_K}|X = \boldsymbol{x}) > \cdots > P(Y = m_{i_1}|X = \boldsymbol{x}). \tag{7}$$

We can then pick out one of the undominated labels, i.e., one with ***maximal probability***.

Example 1. *Given a set of labels $\mathcal{K} = \{m_a, m_b, m_c\}$, a new unlabeled instance $\boldsymbol{x}$, and the probability estimates of the conditional distribution $\widehat{\mathbb{P}}_{\boldsymbol{Y}|X}$:*

$$\widehat{P}(Y = m_a|X = \boldsymbol{x}) = 0.3$$
$$\widehat{P}(Y = m_b|X = \boldsymbol{x}) = 0.1$$
$$\widehat{P}(Y = m_c|X = \boldsymbol{x}) = 0.6$$

the complete preorder between labels w.r.t estimated probabilities is $m_c \succ m_a \succ m_b$.

$\{m_c\}$ being the maximal label dominating other ones (Figure 2), it is the predicted one.

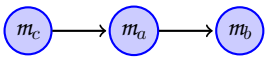
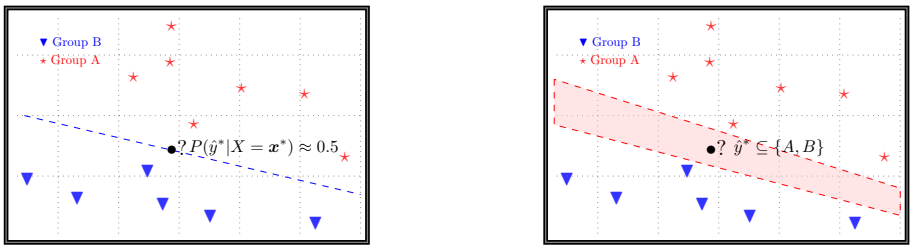


Figure 2: Graph of complete preorder on labels

In the case where we find $1 < r \leq K$ equal maximal conditional probabilities, which is unlikely in practice but not impossible, they can be considered as *indifferent* and chosen randomly. It should be noted that, whatever the quantity of data used to induce the model or the specific new instance $\boldsymbol{x}$ we face (that may come from a poorly populated region), we will always get (up to indifference) a unique undominated model. In contrast, the IP approach where we consider sets $\mathscr{P}_{Y|X}$ may result in partial orders having multiple undominated and incomparable labels.

2.2. Classification with imprecise probabilities

Often, the decision maker can be faced with unreliable or hard situations where making a *single* decision may give rise to serious mistakes (e.g. cancer screening). The hardness of such situations can for instance be due to the lack of sufficient evidence or information (i.e. *uncertainty* in data). Such cases could be dealt more reliably via a cautious decision (i.e. a set of plausible choices, see e.g., Figure 3).



(a) Precise decision-making

(b) Cautious decisions-making

Figure 3: *Cautious vs precise decision-making.* Figure (a) shows a precise model, where there are no regions where the model will output set-valued predictions, in contrast with (b) where such a region exists (in red).

As shown by Equation 7 and Example 1, usual statistics and probabilities usually model *uncertainty* in data with single distribution $\mathbb{P}$, canonically ending up in a unique undominated label. While it is possible to implement decision rules providing set-valued predictions in such settings [10], several authors [14, 22] have

argued that a single distribution cannot always faithfully represent lack of information. Following them, we consider the framework of imprecise probabilities to account for such lack of information.

In what follows, we will introduce some essential concepts about imprecise probabilities, as well as an extension of the *precise* decision approach given in Definition 2 relying on the convex set of distributions $\mathscr{P}_{y|X}$. As in this work we want to obtain more reliable decision by allowing partial predictions, i.e. cautious decisions, we will focus on extensions where we get a set-value label $\widehat{Y}$ instead of a *precise* label $\widehat{y}$.

2.2.1. Basic notions about imprecise probabilities

Imprecise probability theory often (and will in our case) consists in representing our uncertainty by a convex set $\mathscr{P}_X$ of probability distributions [14, 7] (i.e. a *credal set* [19]), defined over a space $\mathscr{X}$ rather than by a precise probability measure $\mathbb{P}_X$ [23]. As they include precise distributions as special cases, such convex sets of distributions provides richer, more expressive models of uncertainty that allow us to better describe uncertainty originating from imperfect or scarce data.

Given such a set of distribution $\mathscr{P}_X$ and any measurable event $A \subseteq \mathscr{X}$, we can define the notions of lower and upper probabilities $\underline{P}_X(A)$ and $\overline{P}_X(A)$, respectively as:

$$\underline{P}_X(A) = \inf_{P \in \mathscr{P}_X} P(A) \qquad \text{and} \qquad \overline{P}_X(A) = \sup_{P \in \mathscr{P}_X} P(A) \tag{8}$$

where $\underline{P}_X(A) = \overline{P}_X(A)$ only when we have sufficient information about A.

Estimations of parameters in the context of imprecise probabilities is usually more complicated as we consider a set $\mathscr{P}_X$ of distributions instead of a *single* distribution $\mathbb{P}_X$. In our case, such complications will be limited, as we will rely on previous works providing efficient generalized Bayesian inference methods for exponential families (which include Gaussian distributions), that we will present in Section 3.2 For theoretical developments of the next subsection about decision, we will assume that we know the set $\mathscr{P}_{Y|X}$ of conditional distributions over the classes.

2.2.2. Decision making under imprecise probabilities

Within IP theories, we can find different methods extending the decision criterion given in Definition 2 (more details in [21]). For classifying a new instance $\boldsymbol{x}$, we will make use of the *maximality criterion* [7, §8.6] that has strong theoretical justifications [14, §3.9.5] and often remains applicable in practice [24, 25, 26]. This one extends Equation (5) in a robust way, requiring that a preference holds only if it holds for every model. More precisely, the criterion of maximality is defined as follows:

Definition 3 (Partial Ordering by Maximality Criterion [21, §3.2])**.** *Let $\mathcal{L}(\cdot,\cdot)$ be a general loss function and $\mathscr{P}_{Y|X}$ a set of probability distributions, then under the maximality criterion, m_a is preferred to m_b iff the cost of exchanging m_a with m_b have a positive lower expectation:*

$$m_a \succ_M m_b \iff \inf_{\mathbb{P}_{Y|\boldsymbol{x}} \in \mathscr{P}_{Y|\boldsymbol{x}}} \mathbb{E}_{\mathbb{P}_{Y|\boldsymbol{x}}}[\mathcal{L}(\cdot, m_b) - \mathcal{L}(\cdot, m_a)|\boldsymbol{x}] > 0 \tag{9}$$

if $\mathcal{L}(\cdot,\cdot)$ is 0/1 loss function, this gives:

$$m_a \succ_M m_b \iff \inf_{\mathbb{P}_{Y|\boldsymbol{x}} \in \mathscr{P}_{Y|\boldsymbol{x}}} \Big[P(Y = m_a|X = \boldsymbol{x}) - P(Y = m_b|X = \boldsymbol{x})\Big] > 0 \tag{10}$$

Equation (10) amounts to asking that Equation (5) is true for all possible probability distributions in $\mathscr{P}_{Y|\boldsymbol{x}}$. In practice, $\succ_M$ can be a partial order with several maximal elements, in which case the prediction becomes imprecise due to high uncertainty in the model. Note that when $N \to \infty$, imprecise and precise models will usually coincide. The prediction $\widehat{Y}_M$ resulting from $\succ_M$ is defined as:

$$\widehat{Y}_M = \Big\{ m_a \in \mathscr{Y} \;\Big|\; \nexists m_b \in \mathscr{Y} : m_b \succ_M m_a \Big\} \tag{11}$$

Example 2. *Given the label set* $\mathcal{K} = \{m_a, m_b, m_c\}$*, we could have the following plausible partial ordering:*

$$\mathscr{B} = \{m_a \succ_M m_b, m_c \succ_M m_b\}$$

where $\widehat{Y}_M = \{m_a, m_c\}$ *is the predicted set obtained from set* $\mathscr{B}$ *of comparisons by the criterion of maximality (figure 4).*

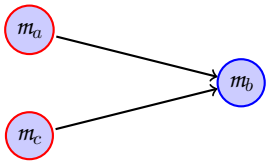


Figure 4: Graph of partial order of set $\mathscr{B}$.

In the next section, we expose how the Gaussian discriminant model can be made imprecise to characterise the conditional distributions $\mathbb{P}_{Y|X}$ by a set $\mathscr{P}_{Y|X}$. We will first recall the precise model, before making it imprecise.

3. Gaussian discriminant analysis model

As mentioned in the introduction, a classical way to estimate the distribution $\mathbb{P}_{Y|X}$ is by using Bayes' theorem. Making use of Equation (1), we will discuss the precise and imprecise approach, respectively in Sections 3.1 and 3.2.

3.1. Statistical inference with precise probabilities

There are many ways to model $\mathbb{P}_{X|Y=m_k}$, but in this work, we focus on *parametric* discriminant analysis which assume that $\mathbb{P}_{X|Y=m_k}$ follows a multivariate Gaussian distribution $\mathcal{N}(\mu_{m_k}, \Sigma_{m_k})$ with unknown mean μ_{m_k} and covariance matrix Σ_{m_k}, i.e.:

$$\mathscr{G}_{m_k} := \mathbb{P}_{X|Y=m_k} \sim \mathcal{N}(\mu_{m_k}, \Sigma_{m_k}) \tag{12}$$

whose probability density function is written

$$P(X = \boldsymbol{x}|Y = m_k) = \frac{1}{(2\pi)^{p/2}|\Sigma_{m_k}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\mu_{m_k})^T \Sigma_{m_k}^{-1}(\boldsymbol{x}-\mu_{m_k})}. \tag{13}$$

The marginal distribution is defined as a multinomial $\pi_y := \mathbb{P}_Y$, where $P(Y = m_k) = \pi_{m_k}$. So, under a 0/1 loss function, the optimal prediction becomes:

$$\widehat{\varphi}(\boldsymbol{x}|\theta_{m_k}) := \arg\max_{m_k \in \mathcal{K}} \ \log \pi_{m_k} - \frac{1}{2}\log|\Sigma_{m_k}| - \frac{1}{2}(\boldsymbol{x}^T - \mu_{m_k})^T \Sigma_{m_k}^{-1}(\boldsymbol{x}^T - \mu_{m_k}) \tag{14}$$

where $\Theta = \{\theta_{m_k} | \theta_{m_k} = (\pi_{m_k}, \Sigma_{m_k}, \mu_{m_k}), \forall m_k \in \mathcal{K}\}$ is the parametric space from which comes our estimate. In Table 2, we remind different discriminant models arising from of the last equation, and corresponding to various constraints imposed to the covariance matrices of the conditional distributions.

Discriminant analysis model	Assumptions ($\forall m_k \in \mathcal{K}$)	Parametric space ($\forall m_k \in \mathcal{K}$)
Parametric Gaussian conditional distribution $\mathbb{P}_{y\|X}$		
Linear Discriminant [6, §4.3]	Homoscedasticity: $\Sigma_{m_k} = \Sigma$	$\Theta = \{\theta_{m_k} \| \theta_{m_k} = (\pi_{m_k}, \Sigma, \mu_{m_k})\}$
Quadratic Discriminant [6, §4.3]	Heteroscedasticity: $\Sigma_{m_k} = \Sigma_k$	$\Theta = \{\theta_{m_k} \| \theta_{m_k} = (\pi_{m_k}, \Sigma_k, \mu_{m_k})\}$
Naive Discriminant [6, §6.63]	Feature independence: $\Sigma_{m_k} = \boldsymbol{\sigma}_k^T \mathbb{I}$	$\Theta = \{\theta_{m_k} \| \theta_{m_k} = (\pi_{m_k}, \sigma_k, \mu_{m_k})\}$
Euclidean Discriminant [27]	Unit-variance feature indep.: $\Sigma_{m_k} = \mathbb{I}$	$\Theta = \{\theta_{m_k} \| \theta_{m_k} = (\pi_{m_k}, \mu_{m_k})\}$

Table 2: Gaussian discriminant analysis models

In frequentist inference, usual estimation of parameters of (14) is obtained by MLE using a subset $\mathscr{D}_{m_k} = \{(x_{i,k}, y_{i,k=m_k}) | i = 1, \ldots, n_k\} \subseteq \mathscr{D}$ of observations of training data. We have $\hat{\pi}_{m_k} = n_k/N$ (frequency

of m_k) and $\hat{\mu}_{m_k} = \overline{\boldsymbol{x}}_k$ (sample mean of $\mathscr{D}_{m_k}$). Depending on whether we assume the model to have (1) dependent features, we will have an hetero- or homo-scedastic assumption, with respectively $\hat{\Sigma}_{m_k} = \widehat{S}_{m_k}$ (sample covariance matrix of $\mathscr{D}_{m_k}$) or $\hat{\Sigma}_{m_k} = \widehat{S}$ (within-class covariance matrix $\mathscr{D}$), or to have (2) independent features, we will have features weighted proportionally to their inverse variance $\hat{\Sigma}_{m_k} = \hat{\boldsymbol{\sigma}}_k^T \mathbb{I}$ or unweighed with all weights equal to 1, i.e. $\hat{\Sigma}_{m_k} = \mathbb{I}$.

Those estimates do not account for the quantity of data they are based on, which may be low to start with, and may also vary significantly across classes, especially in case of imbalanced data sets. To solve this issue, we propose in the next section an imprecise discriminant model, based on the use of imprecise probabilities and using results from Benavoli *et al.* [18].

3.2. Statistical inference with imprecise probabilities

To estimate $\mathbb{P}_{X|y}$ and $\mathbb{P}_Y$ in the form of convex sets of distributions, we will use robust Bayesian inference under prior near-ignorance models. Before describing our *imprecise* estimation, we make three general assumptions for our *imprecise* Gaussian discriminant model:

1. Normality of conditional probability distribution $\mathbb{P}_{X|Y=m_k} := \mathscr{G}_{m_k}$, as in the classical case.
2. A *precise* estimation of marginal distribution $\mathbb{P}_Y := \widehat{\pi}_y$.
3. A *precise* estimation of covariance matrix $\Sigma_k := \widehat{\Sigma}_k = \widehat{S}_k$ or $\widehat{S}$.

In Section 6, we will discuss the relaxation of assumption 2, considering a set of distributions $\mathscr{P}_Y$.

3.2.1. Robust Bayesian inference

The estimation of parameters in Bayesian inference relies mainly on two components; the *likelihood function* and the *prior distribution*, from which posterior inferences can then be made on unknown parameters of the model, in our case θ_{m_k}.

In the particular case of $\mathbb{P}_{X|Y=m_k}$, the *likelihood function* is the product of conditional probabilities $\prod_i^{n_k} P_{x_{i,k}|y_{i,k},\theta_{m_k}}$ and the *prior distribution* $\mathbb{P}_{\theta_{m_k}}$ models our knowledge about $\theta_{m_k} = (\Sigma_{m_k}, \mu_{m_k})$. In this paper, we focus on estimating imprecise ***mean parameters*** (i.e. $\theta_{m_k} = \mu_{m_k}$), assuming a (precise) estimation of $\hat{\Sigma}_{m_k}$, for reasons of computational complexity that will be discussed in Section 7. Thus, the posterior on the mean is such that

$$P(\mu_{m_k} \mid \mathscr{D}_{m_k}) \propto \prod_i^{n_k} P(X = \boldsymbol{x}_{i,k} \mid \mu_{m_k}, y_{i,k} = m_k) P(\mu_{m_k}). \tag{15}$$

To simplify, we will from now on remove the subscript m_k, always bearing in mind that these estimations are related to a group of observations labelled m_k.

3.2.2. Near-ignorance on Gaussian discriminant analysis

Near-ignorance models allow us to provide an *"objective inference"* approach, representing *ignorance about unknown parameter* and *letting the data speak for themselves*. In their work, Benavoli *et al* in [18] propose a new near-ignorance model based on a set of distribution $\mathscr{M}$, which aims to reconcile two approaches, namely, the re-parametrization invariance and the *Walley's* near-ignorance prior. For that, *Benavoli A. et al* define four minimal properties, which must be satisfied whenever there is no prior information about the unknown parameter, on the set of distributions $\mathscr{M}$ (more details in [18, §2]).

(P1) **Prior-invariance,** that states that $\mathscr{M}$ should be invariant under some re-parametrization of the parameter space (e.g. translation, scale, permutation, symmetry, etc).

(P2) **Prior-ignorance,** that states that $\mathscr{M}$ should be sufficiently large for reflecting a complete absence of prior information w.r.t unknown parameter, but no too large to be incompatible with property **(P3)**.

(P3) **Learning from data,** that states that $\mathscr{M}$ should always provide non-vacuous posterior inferences, in other words, it should learn from the observations.

(P4) **Convergence,** that states that the influence of $\mathscr{M}$ on the posterior inference vanishes when increasing number of observations, i.e. $n \to \infty$, requiring consistency with the precise approach at limit.

Benavoli *et al* [18] provide a set of conjugate priors $\mathscr{M}$ for *regular multivariate exponential families* [28, §3.3.4] ($\mathscr{FE}\mathscr{xp}$) that satisfies the last four properties under quite weak assumptions. Borrowing from [18], we can define this set of prior distribution $\mathscr{M}$ as follows:

Definition 4 (Prior near-ignorance for k-parameter exponential families [18, §4, eq. 16])**.** *Let* $\mathbb{L}$ *be a bounded closed convex subset of* $\mathbb{R}^k$ *strictly including the origin ([18, lem. 4.5]).*

$$\mathbb{L} = \left\{\ell \in \mathbb{R}^k : \ell_i \in [-c_i, c_i], c_i > 0, i = \{1, \ldots, d\}\right\} \tag{16}$$

Let $w \in \mathscr{W} = \mathbb{R}^k$ *be a real-valued parameter with a density having the following functional form that belongs to* $\mathscr{FE}\mathscr{xp}$*:*

$$p(w) = \frac{\ell}{\exp(\ell^T r)} exp(\ell^T w) \mathbb{1}_{\mathscr{W}_r}(w) \tag{17}$$

where ℓ *belongs to* $\mathbb{L}$ *and* $r \in \mathbb{R}^k$ *is a real value. The set of prior distributions (c.f. [18, th. 4.6]) can be written as follows:*

$$\mathscr{M}^w = \left\{w \in \mathscr{W} \mid p(w) \propto \exp(\ell^T w), \ell = [\ell_1, \ldots, \ell_k] \in \mathbb{L}\right\} \tag{18}$$

Since our Gaussian probability distribution $\mathbb{P}_{X|y=m_k}$ given by Equation (12) belongs to $\mathscr{FE}\mathscr{xp}$, we can use the set of prior distributions $\mathscr{M}^\mu$ of Equation (18) in order to get a set of posterior distributions $\mathscr{M}_n^\mu$ having the same functional form ($\mathscr{FE}\mathscr{xp}$) [17, §5.2]:

$$\mathscr{M}_n^\mu = \left\{ \mu \middle| \overline{\boldsymbol{x}}_n, \ell \propto \mathcal{N}\left(\frac{\ell^T \widehat{\Sigma} + n\overline{\boldsymbol{x}}_n}{n}, \frac{1}{n}\widehat{\Sigma}\right) \middle| \begin{array}{l} \mu \in \mathbb{R}^n, \\ \ell \in \mathbb{L} \end{array} \right\} \tag{19}$$

where $\overline{\boldsymbol{x}}_n = \frac{1}{n}\sum_{i=1}^n x_i$ and $\ell \in \mathbb{L}$. We can then estimate the lower and upper values of the unknown μ parameters, so for every dimension $i \in \{1, \ldots, d\}$ [18]:

$$\inf_{\mathscr{M}_n^\mu} \mathbb{E}[\mu_i \mid \overline{\boldsymbol{x}}_n, \ell] = \underline{\mathbb{E}}[\mu_i \mid \overline{\boldsymbol{x}}_n, \ell] = \frac{-c_i + n\overline{\boldsymbol{x}}_n}{n} \tag{20}$$

$$\sup_{\mathscr{M}_n^\mu} \mathbb{E}[\mu_i \mid \overline{\boldsymbol{x}}_n, \ell] = \overline{\mathbb{E}}[\mu_i \mid \overline{\boldsymbol{x}}_n, \ell] = \frac{c_i + n\overline{\boldsymbol{x}}_n}{n} \tag{21}$$

As a result, we will have for each label m_k a convex space of plausible values for the mean μ_{m_k} which can be represented by the hyper-cube

$$\mathbb{G}_{m_k} = \left\{ \widehat{\mu}_{m_k} \in \mathbb{R}^d \middle| \widehat{\mu}_{i,m_k} \in \left[\frac{-c_i + n_k \overline{\boldsymbol{x}}_{i,n_k}}{n_k}, \frac{c_i + n_k \overline{\boldsymbol{x}}_{i,n_k}}{n_k}\right], \forall i = \{1, ..., d\} \right\}. \tag{22}$$

Remark 1. *The convergence property (P4) ensures us that no matter the initial value of our convex space* $\mathbb{L}$*, when the number of observations tends to infinity,* $n \to \infty$*, their influence on the posterior inference of* $\hat{\mu}$ *will disappear, i.e* $\mathbb{G}_{m_k} \xrightarrow[n\to\infty]{} \overline{\boldsymbol{x}}_n$*, and will become the asymptotic estimator of the precise Gaussian distribution.*

On the basis of the set $\mathbb{G}_{m_k}$ previously calculated, we can simply consider the following set of conditional probability distributions $\mathscr{P}_{X|y=m_k}$ (or set of predictive distributions) for every label m_k on $\mathscr{K}$:

$$\mathscr{P}_{X|y=m_k} = \left\{ \mathbb{P}_{X|Y=m_k} \middle| \mathbb{P}_{X|Y=m_k} \sim \mathcal{N}(\mu_{m_k}, \widehat{\Sigma}_{m_k}), \mu_{m_k} \in \mathbb{G}_{m_k} \right\} \tag{23}$$

In what follows, we study how we can incorporate the sets of distributions $\mathscr{P}_{X|Y=m_k}$ in Gaussian discriminant analysis, using maximality (Definition 3) to get our (possibly) *imprecise* classification.

4. Imprecise Classification with $\mathscr{L}_{0/1}$ loss function

Let us now present our approach to make cautious classification by using sets of conditional distribution given by Equation (23) and obtained from a near-ignorance model. Using the maximality criterion, to know whether $m_a \succ_M m_b$, we need to solve Equation (10) by applying Bayes' theorem:

$$\inf_{\mathbb{P}_Y \in \mathscr{P}_Y} \inf_{\substack{\mathbb{P}_{X|m_a} \in \mathscr{P}_{X|m_a} \\ \mathbb{P}_{X|m_b} \in \mathscr{P}_{X|m_b}}} P(X = \boldsymbol{x}|Y = m_a)P(Y = m_a) - P(X = \boldsymbol{x}|Y = m_b)P(Y = m_b) > 0 \tag{24}$$

since the marginal $P(X = \boldsymbol{x}) = \sum_{m_l \in \mathscr{Y}} P(X = \boldsymbol{x}|Y = m_l)P(Y = m_l)$, which is the same positive constant of normalisation for each probability, can be omitted.

As conditional distributions sets $\mathscr{P}_{X|Y=m_k}$ are independent of each others, we can rewrite Equation (24) as follows (cf. [24, eq. 4.3]):

$$\inf_{\mathbb{P}_Y \in \mathscr{P}_Y} \underline{P}(X = \boldsymbol{x}|Y = m_a)P(Y = m_a) - \overline{P}(X = \boldsymbol{x}|Y = m_b)P(Y = m_b) > 0 \tag{25}$$

where $\underline{P}$ ($\overline{P}$) is the infimum (supremum) conditional probability. Also, applying Assumption 2 and the fact that every $\widehat{\pi}_y > 0$, solving Equation (25) is reduced to finding the two values

$$\underline{P}(X = \boldsymbol{x}|Y = m_a) = \inf_{\mathbb{P}_{X|m_a} \in \mathscr{P}_{X|m_a}} P(X = \boldsymbol{x}|Y = m_a), \tag{26}$$

$$\overline{P}(X = \boldsymbol{x}|Y = m_b) = \sup_{\mathbb{P}_{X|m_b} \in \mathscr{P}_{X|m_b}} P(X = \boldsymbol{x}|Y = m_b) \tag{27}$$

As $\mathscr{P}_{X|y=m_k}$ is a set of Gaussian distributions, the solutions of Equations (26) and (27) are respectively obtained for the following values of the means

$$\underline{\mu}_{m_a} = \arg\inf_{\mu_{m_a} \in \mathbb{G}_{m_a}} -\frac{1}{2}(\boldsymbol{x} - \mu_{m_a})^T \widehat{\Sigma}_{m_b}^{-1}(\boldsymbol{x} - \mu_{m_a}), \tag{28}$$

$$\overline{\mu}_{m_b} = \arg\sup_{\mu_{m_b} \in \mathbb{G}_{m_b}} -\frac{1}{2}(\boldsymbol{x} - \mu_{m_b})^T \widehat{\Sigma}_{m_b}^{-1}(\boldsymbol{x} - \mu_{m_b}), \tag{29}$$

where $\widehat{\Sigma}_{m_b}^{-1}$ is the inverse of the covariance matrix (Assumption 3). Depending on the internal structure of the *precise* covariance matrix $\hat{\Sigma}_k$, solving for (28) and (29) may be more or less computationally challenging. We will consider two main different imprecise discriminant models: (1) with non-diagonal covariance matrix and (2) with diagonal covariance matrix.

4.1. Gaussian discriminant model with dependent features

Similarly to the distinction made in the precise case, we will consider two different variants of the non-diagonal case.

Case 1. *Imprecise Quadratic discriminant analysis (IQDA): if we suppose that the covariance structures of all groups of observations are different, that is* $\widehat{\Sigma}_{m_k} = \widehat{S}_{m_k}, \forall m_k \in \mathscr{K}$.

Case 2. *Imprecise linear discriminant analysis (ILDA): if we assume that all groups of observations have the same covariance structure, that is* $\hat{\Sigma}_{m_k} = \hat{S}, \forall m_k \in \mathscr{K}$.

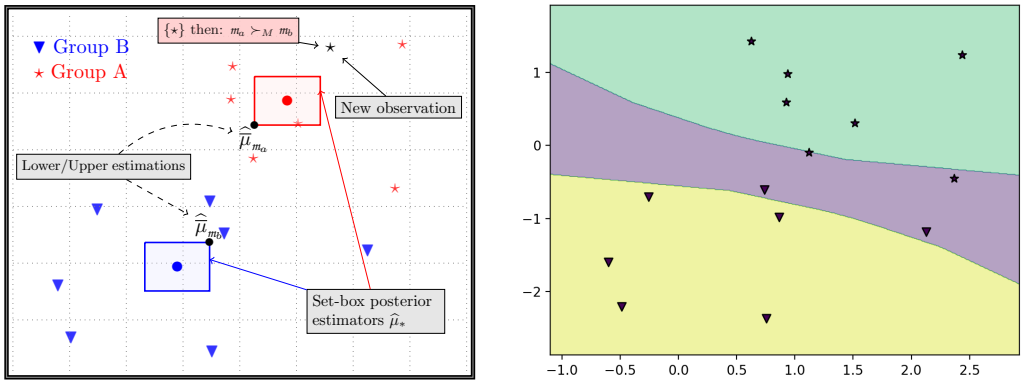
In those cases where the covariance matrix contains collinear columns, $\hat{\Sigma}_{m_k}$ will not be invertible, in which case we use the *singular value decomposition* (SVD) method for computing the *pseudo-inverse* of covariance matrix. Before studying the computational issues of IQDA and ILDA, i.e. Equations (28) and (29), we will illustrate the last case (ILDA) in Example 3.

Example 3. *The interest of modelling an imprecise mean is to be able to detect areas where we should be cautious and predict sets of labels rather than a single one. For example, in Figure 5, we simulated two groups of observations $x_{m_a *}$ and $x_{m_b *}$ (i.e. binary case), each with two non-correlated regressors and different means:*

$$\begin{pmatrix} x_{m_a 1} \\ x_{m_a 2} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0.25 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$
$$\begin{pmatrix} x_{m_b 1} \\ x_{m_b 2} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0.5 \\ -1.0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$
$$\mathbb{L} = \left\{\ell \in \mathbb{R}^2 : \ell_i \in [-c_i, c_i], c_i = 2\right\}$$

Figure 5(a) illustrates this example and pictures the following things: groups of observations $x_{m_a *}$ and $x_{m_b *}$ with the symbols $\star$ and $\blacktriangledown$, respectively, and the posterior convex estimates $\mathbb{G}$ (solid) of the means after injecting the information contained in the training data.

We also drew the (precise) mean of each group, i.e. μ_{m_a} and μ_{m_b}, as solid points, and a black dot ($\bullet$) representing a new unlabelled instance $\boldsymbol{x}$ as well as positions of solutions of Equations (28) and (29). In Figure 5, we observe (in purple) an area of uncertainty generated by the imprecise mean and the maximality criterion.



(a) Upper/lower estimation of μ_k (b) Imprecise decision area

Figure 5: *Imprecise boundary area and estimation.* Figure 5(a) shows an example of imprecise estimation of means μ_*, and Figure 5(b) shows an imprecise decision area of purple colour where the subset $\widehat{Y}=\{m_a, m_b\}$ of labels is the imprecise decision, that is in this region m_a and m_b are incomparable.

Let us now discuss the problem of solving Equations (28) and (29). Expressing $\mathbb{G}_{m_b}$ as constraints, the solution $\overline{\mu}_{m_b}$ of (29) can be written as

$$\begin{aligned} \overline{\mu}_{m_b} &= \arg\sup \; -\frac{1}{2}\hat{\mu}_{m_b}^T \hat{\Sigma}_{m_b}^{-1} \hat{\mu}_{m_b} + q^T \hat{\mu}_{m_b} \\ s.t. \; &\frac{-c_j + n_{m_b}\overline{x}_{j,n}}{n_{m_b}} \leq \hat{\mu}_{j,m_b} \leq \frac{c_i + n_{m_b}\overline{x}_{j,n_{m_b}}}{n_{m_b}}, \\ &q^T = -\boldsymbol{x}^T \hat{\Sigma}_{m_b}^{-1}, \quad \forall j = \{1, ..., d\} \end{aligned} \tag{BQP}$$

This optimisation problem is well-known as a box-constraint quadratic program (BQP) [29], as (1) the constraint space $\mathbb{G}_{m_k}$ is a convex space, and (2) $\widehat{\Sigma}_{m_k}^{-1}$ is a positive (semi)-definite matrix, pending the fact that the covariance matrix $\hat{\Sigma}_{m_k}$ does not have multicollinearity problems [30]. Computing an optimal global solution of (BQP) in polynomial time is easy using modern optimisation libraries (e.g. using the CvxOpt python library [31]), as we have to maximize a concave function (or, equivalently, minimise a convex one).

Finding $\underline{\mu}_{m_a}$ in equation (28) is much more difficult, as one seeks to solve the optimization problem

$$\underline{\mu}_{m_a} = \arg\inf_{\hat{\mu}_{m_a} \in \mathbb{G}_{m_a}} \; -\frac{1}{2}\mu_{m_a}^T \hat{\Sigma}_{m_a}^{-1} \hat{\mu}_{m_a} + q^T \hat{\mu}_{m_a} \tag{NBQP}$$

That comes down this time to maximizing a convex function over box-constraints ($\mathbb{G}_{m_a}$), which is known to be NP-Hard [32]. To solve it, we use a brand-and-bound (B&B) algorithm [33, 34], that employs a finite branching based on the first-order Karush-Kuhn-Tucker[5] conditions and polyhedral semidefinite relaxation in each node of the B&B tree (more details in [33]).

4.2. Gaussian discriminant model with independence features

When the number of features becomes high, and the associated optimisation problem quite time-consuming to solve, it may be interesting to consider some additional assumptions which will significantly reduce the inference complexity. In what follows, we will assume that features x_i^j are independent conditional on the label m_k.This translates in the fact that covariance matrices become diagonal matrices, i.e. $\Sigma_{m_k} = \boldsymbol{\sigma}_{m_k}^T \mathbb{I}$ with $\boldsymbol{\sigma}_{m_k}^T = (\sigma_{m_k}^1, \ldots, \sigma_{m_k}^p)$ a p-dimensional vector containing the variance of each feature, which can be interepreted as weights of the features. Therefore, we can rewrite Equations (28) and (29) as follows:

$$\underline{\mu}_{m_a} = \arg \inf_{\mu_{m_a} \in \mathbb{G}_{m_a}} -\frac{1}{2} \boldsymbol{w}_{m_a} \left\| \boldsymbol{x} - \mu_{m_a} \right\|^2 \tag{30}$$

$$\overline{\mu}_{m_b} = \arg \sup_{\mu_{m_b} \in \mathbb{G}_{m_b}} -\frac{1}{2} \boldsymbol{w}_{m_b} \left\| \boldsymbol{x} - \mu_{m_b} \right\|^2 \tag{31}$$

where $\boldsymbol{w}_{m_k} = (w_{m_k}^1, \ldots, w_{m_k}^p)^T$ such that $w_{m_k}^j = 1/\sigma_{m_k}^j, \forall j \in \{1, \ldots, p\}$, in this scenario, we will consider two new models.

Case 3. *Imprecise naive discriminant analysis (INDA): this case is similar to the Naive Bayes classifier, as we simply consider the assumption $\Sigma_{m_k} = \widehat{\boldsymbol{\sigma}}_k^T \mathbb{I}$ where $\widehat{\boldsymbol{\sigma}}_k$ are the empirical variance estimator obtained from a group of observation belonging to the label m_k.*

Case 4. *Imprecise Euclidian discriminant analysis (IEDA): this is a case more specific than INDA, where we assume that for every $j \in \{1, \ldots, p\}$ we have $\widehat{\sigma}_k^j = 1$, meaning that the measure used to evaluate the probability of a label given a new instance is proportional to the Euclidian distance between the instance and the corresponding mean. The Euclidean classifier is one of the simplest existing classifier, and is the supervised counterpart of the standard k-means method.*

We show below that when the covariance matrix is diagonal, optimisation problems (30) and (31) become very easy (i.e. linear in p, $\mathcal{O}(p)$) to solve.

Proposition 1. *For two vectors $\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^p$, and a box-convex space on $\mathbb{R}^p$:*

$$\mathbb{G} = \left\{ \boldsymbol{\mu} \in \mathbb{R}^p \;\middle|\; \mu^j \in [\underline{\mu}^j, \overline{\mu}^j], \forall j \in \{1, \ldots, p\} \right\}$$

- the infimum weighted distance subject to constraints $\mathbb{G}$ is:

$$\inf_{\mu \in \mathbb{G}} -\frac{1}{2} \boldsymbol{w}^T \left\| \boldsymbol{x} - \mu \right\|^2 = -\frac{1}{2} \sum_j^p w^j \max_j \{(x^j - \underline{\mu}^j)^2, (x^j - \overline{\mu}^j)^2\} \tag{32}$$

- and the supremum weighted distance subject to same constraints is:

$$\sup_{\mu \in \mathbb{G}} -\frac{1}{2} \boldsymbol{w}^T \left\| \boldsymbol{x} - \mu \right\|^2 = -\frac{1}{2} \sum_j^p w^j \begin{cases} 0 & \text{if } x^j \in [\underline{\mu}^j, \overline{\mu}^j] \\ \min_j \{(x^j - \underline{\mu}^j)^2, (x^j - \overline{\mu}^j)^2\} & \text{otherwise} \end{cases} \tag{33}$$

[5] Also known as KKT, which allows to solve problems of optimisation subject to non-linear constraints in the form of inequalities.

Proof. Since each element of the sum is positive, we can interchange the infimum operator with summation, and calculate the supremum of each component as follows:

$$\inf_{\mu\in\mathbb{G}} -\frac{1}{2}\sum_{j}^{p} w^j(x^j-\mu^j)^2 \quad \Longleftrightarrow \quad -\frac{1}{2}\sum_{j}^{p} w^j \sup_{\mu^j\in[\underline{\mu}^j,\overline{\mu}^j]}(x^j-\mu^j)^2 \tag{34}$$

where the supremum can be calculated as follows:

$$\sup_{\mu^j\in[\underline{\mu}^j,\overline{\mu}^j]}(x^j-\mu^j)^2 = \max_j\{(x^j-\underline{\mu}^j)^2,(x^j-\overline{\mu}^j)^2\} \tag{35}$$

In the second case and for similar reasons, we can also put the supremum operator inside of summation and calculate of infimum value of each component:

$$\sup_{\mu\in\mathbb{G}} -\frac{1}{2}\sum_{j}^{p} w^j(x^j-\mu^j)^2 \quad \Longleftrightarrow \quad -\frac{1}{2}\sum_{j}^{p} w^j \inf_{\mu^j\in[\underline{\mu}^j,\overline{\mu}^j]}(x^j-\mu^j)^2 \tag{36}$$

where the infimum of squared subtraction of each element is:

$$\inf_{\mu^j\in[\underline{\mu}^j,\overline{\mu}^j]}(x^j-\mu^j)^2 = \begin{cases} 0 & \text{if } x^j\in[\underline{\mu}^j,\overline{\mu}^j] \\ \min_j\{(x^j-\underline{\mu}^j)^2,(x^j-\overline{\mu}^j)^2\} & \text{otherwise} \end{cases} \tag{37}$$

□

The next section presents some experiences with different data sets and different precise and imprecise models.

5. Experiments setting

In this section, we provide experimental results evaluating the performance of our different imprecise Gaussian discriminant models (cf. Section 4).

5.1. How can we choose parameter c_i?

The choice of parameters c_i determines the amount of imprecision in our posterior inference. It should be large enough to guarantee more reliable predictions when missing information, but small enough so as to provide informative predictions when possible. Therefore, in the absence of prior information and for symmetry reasons, we will consider a symmetric box around 0, as follows:

$$\mathbb{L}' = \left\{\ell\in\mathbb{R}^k : \ell_i\in[-c,c], c>0, i=\{1,\ldots,d\}\right\}. \tag{38}$$

In order to fix a value of c, there exists different approaches already mentioned in Section 4.3 of [18]. One can for example rely on the rate of convergence of the lower and upper posterior expectations [14]:

$$\forall i \quad \left(\overline{E}[\mu_i|\overline{\boldsymbol{x}}_n,\ell]-\underline{E}[\mu_i|\overline{\boldsymbol{x}}_n,\ell]\right) = \frac{2c}{n} \xrightarrow[n\to\infty]{} 0 \tag{39}$$

meaning that for small values of c, we would reach a faster convergence of Equation (39) to a *precise* posterior inference (as precise models). A value of $c\leq 0.75$ is recommended by [18, §4.3, §8], however since we are in a classification problem, we will select an optimal value of c through cross-validation on the training samples. More precisely, we restrict c to the interval $[0.01, 5]$, discretised into $[0.01, 0.02, ..., 5]$, with the optimal value decided by cross validation on the training samples. A typical empirical evolution of the accuracy measures used in the next sections is shown in Figure 8 for the four IGDA methods. It clearly shows that performances first increase in average with imprecision, but then degrades as imprecision becomes too large.

#	name	# instances	# features	# labels
a	iris	150	4	3
b	wine	178	13	3
c	forest	198	27	4
d	seeds	210	7	3
e	glass	214	9	6
f	ecoli	336	7	8
g	libras	360	91	15
h	dermatology	385	34	6
i	vehicle	846	18	4
j	vowel	990	10	11
k	yeast	1484	8	12
l	wine quality	1599	11	6
m	optdigits	1797	64	10
n	segment	2300	19	7
o	wall-following	5456	24	4

Table 3: Data sets used in the experiments

5.2. Data sets and experimental setting

We perform experiments on 15 data sets issued from UCI machine repository [35](cf. Table 3), following a 10×10-fold cross-validation procedure. We aim to compare the performance of our imprecise Gaussian classifier model approach with the existing precise models (c.f. Table 2).

Owing to small amounts of samples in some groups of observations (belonging to a specific label m_k) of some data sets, the QDA model can suffer from a phenomenon known as ill-posed covariance matrix (i.e. $n_{m_k} < p$), and in such cases even calculating the pseudo-inverse of $\widehat{\Sigma}_{m_m}$ estimated covariance matrice using SVD method cannot solve the problem. This affects the performance of our classifiers, getting highly significant drop (e.g. in Table 4, glass and yeast data sets). Therefore, in this case specific, we used a basic regularized method for estimated covariance matrice named Regularization QDA (or RQDA)[36, 6]:

$$\Sigma_{m_k}(\alpha) = \alpha\widehat{\Sigma}_{m_k} + (1-\alpha)\mathbb{I}, \tag{40}$$

where $\widehat{\Sigma}_{m_m}$ is the estimated covariance for a group of observations, $\mathbb{I}$ a identity matrix and α the regularization factor.

Comparing indeterminate predictions given in the form of a subset $\widehat{Y}$ of plausible labels against just one determinate prediction $\widehat{y}$ is a hard problem that mostly depends on the circumstances or the context in which a decision-marker may or may not accept partial predictions (or cautious decision) instead of a unique, risky decision. A good evaluation should reward cautiousness provided by $\widehat{Y}$ when it allows to include the true observed label, but not so much as to systematically privilege imprecision over precision. In other words, we need an evaluation metric that seeks a compromise between cautiousness and informativeness. To do this, we adopt the evaluation metric proposed and theoretically justified in [37], called *utility-discounted accuracy*, which makes it possible to reward the imprecision in a more or less strong way. It is written as follows:

$$u(y,Y) = \begin{cases} 0 & \text{si } y \notin Y, \\ \frac{\alpha}{|Y|} - \frac{1-\alpha}{|Y|^2} & \text{autrement.} \end{cases} \tag{41}$$

[37] shows that a value $\alpha = 1$ amounts to not reward cautiousness and to confuse it with randomness, while $\alpha \to \infty$ does not penalize non-informativeness, as the vacuous prediction (i.e. $\widehat{Y} = \mathscr{K}$) would always get a full, guaranteed reward. We will use the usual values u_{65} with $\alpha = 1.6$ and u_{80} with $\alpha = 2.2$ (as in [25]). To have an intuition about these measures, let us simply recall that the u_{65} (u_{80}) measure rewards

a binary correct prediction with 0.65 (0.80), while a purely random, non-cautious guesser picking one of the two possible label would reward it with 0.50. It therefore gives a "reward" of 0.15 (0.30) for rightful cautiousness.

5.3. Experimental results

The average results obtained according to u_{65} and u_{80} utilities, and the average execution time to predict the label of a new unlabeled instance are shown in Table 4.

(a) LDA versus ILDA

	LDA	**ILDA**		**Avg. Time**
#	acc.	u_{80}	u_{65}	
a	97.96 ± 0.05	**98.38**±	97.16±	0.56
b	98.85 ± 0.36	$\mathbf{98.99 \pm 1.17}$	98.95 ± 1.26	1.49
c	94.61 ± 0.60	$\mathbf{94.56 \pm 1.08}$	94.05 ± 1.02	12.14
d	96.35 ± 0.25	$\mathbf{96.59 \pm 0.23}$	96.51 ± 0.23	1.50
e	62.15 ± 0.76	$\mathbf{66.78 \pm 0.73}$	58.87 ± 0.77	
f	87.14 ± 0.37	$\mathbf{88.27 \pm 1.43}$	87.72 ± 1.42	12.40
g	64.45 ± 0.57			
h	96.58 ± 0.35	$\mathbf{97.06 \pm 0.62}$	96.94 ± 0.61	19.24
i	77.96 ± 0.48	$\mathbf{81.98 \pm 0.91}$	79.59 ± 0.82	3.10
j	60.10 ± 0.68	$\mathbf{67.45 \pm 0.48}$	62.41 ± 0.40	4.95
k	58.92 ± 0.17			
l	59.25 ± 0.27	**65.83**±	60.31±	34.85
m	95.40 ± 0.09			
n	$\mathbf{91.60 \pm 0.09}$	90.76 ± 0.35	89.70 ± 0.32	
o	67.96 ± 0.07	**71.34**±	66.65±	10.77
avg.				

(b) QDA versus IQDA

	QDA	**RQDA**	**IQDA**		**Avg. Time**
#	acc.	acc.	u_{80}	u_{65}	
a	97.29 ± 0.44	96.66 ± 4.47	$\mathbf{98.08 \pm 0.41}$	97.13 ± 0.42	
b	99.03 ± 0.45	98.89 ± 2.22	$\mathbf{99.39 \pm 0.14}$	99.09 ± 0.13	
c	89.43 ± 1.34	$\mathbf{97.47 \pm 3.37}$	91.77 ± 1.38	88.90 ± 1.32	
d	94.64 ± 0.47	94.29 ± 2.86	$\mathbf{95.20 \pm 0.26}$	94.72 ± 0.24	
e	7.15 ± 2.39	51.40 ± 9.79	$\mathbf{64.38 \pm 1.36}$	58.36 ± 1.30	
f	46.19 ± 2.97	88.25 ± 5.97	$\mathbf{87.34 \pm 0.90}$	84.79 ± 0.87	
g	34.04 ± 2.14	72.22 ± 6.21			
h	82.47 ± 0.42	96.92 ± 0.88	$\mathbf{84.24 \pm 0.87}$	84.05 ± 0.88	
i	85.07 ± 0.86	85.11 ± 2.63	$\mathbf{87.96 \pm 0.34}$	86.13 ± 0.27	
j	87.83 ± 0.49	87.07 ± 3.49	$\mathbf{89.96 \pm 0.67}$	88.40 ± 0.70	
k	13.18 ± 2.37	56.27 ± 2.29			
l	55.62 ± 0.47	55.79 ± 5.35	**65.85**±	60.36±	
m	87.18 ± 1.31	98.94 ± 0.85			
n	64.69 ± 1.82	91.17 ± 1.61			
o	65.87 ± 0.17	70.56 ± 2.63	$\mathbf{71.79 \pm 0.12}$	69.75 ± 0.12	
avg.					

(c) NDA versus INDA

	NDA	**INDA**		**Avg. Time**
#	acc.	u_{80}	u_{65}	
a	95.07 ± 0.44	95.73 ± 5.78	95.53 ± 5.94	0.46×10^{-3}
b	97.70 ± 0.58	98.39 ± 3.24	93.60 ± 13.93	1.72×10^{-3}
c	95.26 ± 0.33	89.95 ± 15.28	99.95 ± 18.59	4.09×10^{-3}
d	90.38 ± 0.19	91.14 ± 5.88	88.21 ± 11.32	1.12×10^{-3}
e	43.92 ± 1.36	51.25 ± 10.84	50.21 ± 10.41	4.98×10^{-3}
f	82.39 ± 1.22	56.54 ± 18.69	56.54 ± 18.69	3.41×10^{-3}
g	63.17 ± 1.43	65.50 ± 4.76	65.22 ± 5.30	10.35×10^{-3}
h	85.52 ± 0.98	90.71 ± 4.57	90.45 ± 4.64	3.10×10^{-3}
i	45.63 ± 0.89	46.33 ± 7.44	45.70 ± 7.14	1.19×10^{-3}
j	67.26 ± 0.39	70.35 ± 7.44	71.70 ± 10.23	3.14×10^{-3}
k	43.36 ± 0.51	49.51 ± 6.78	50.71 ± 5.81	2.50×10^{-3}
l	54.83 ± 0.34	57.24 ± 8.27	57.33 ± 3.31	1.51×10^{-3}
m	89.69 ± 0.19	85.17 ± 7.75	89.39 ± 7.01	4.44×10^{-3}
n	79.83 ± 0.11	77.11 ± 9.23	78.75 ± 10.31	2.69×10^{-3}
o	52.55 ± 0.12	51.00 ± 6.64	54.32 ± 5.09	1.29×10^{-3}
avg.				

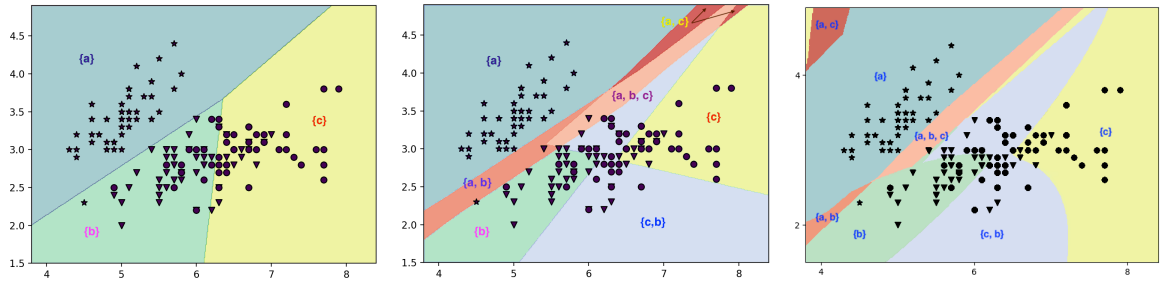
(d) EDA versus IEDA

	EDA	**IEDA**		**Avg. Time**
#	acc.	u_{80}	u_{65}	
a	91.60 ± 0.61	94.80 ± 4.24	93.13 ± 4.75	0.29×10^{-3}
b	46.65 ± 0.85	61.78 ± 3.93	52.04 ± 4.08	0.41×10^{-3}
c	81.09 ± 0.39	82.38 ± 7.42	81.84 ± 9.30	1.31×10^{-3}
d	90.38 ± 0.36	92.86 ± 8.05	87.29 ± 8.94	0.46×10^{-3}
e	46.26 ± 1.68	56.16 ± 12.06	49.08 ± 8.85	0.67×10^{-3}
f	42.59 ± 0.04	43.34 ± 9.40	40.91 ± 7.08	0.87×10^{-3}
g	49.36 ± 1.53	54.62 ± 7.21	49.27 ± 10.53	12.25×10^{-3}
h	51.22 ± 0.92	54.78 ± 10.44	52.87 ± 10.99	1.39×10^{-3}
i	28.03 ± 0.19	45.13 ± 6.03	38.33 ± 4.06	0.67×10^{-3}
j	58.08 ± 0.90	63.94 ± 5.28	58.94 ± 4.60	2.48×10^{-3}
k	31.27 ± 0.13	31.56 ± 3.10	31.39 ± 2.88	2.18×10^{-3}
l	19.72 ± 0.19	22.11 ± 5.64	23.40 ± 8.44	1.03×10^{-3}
m	88.17 ± 0.15	89.11 ± 1.95	88.83 ± 1.82	5.09×10^{-3}
n	22.10 ± 0.11	39.30 ± 5.87	30.65 ± 1.10	2.15×10^{-3}
o	57.90 ± 0.11	56.66 ± 7.01	58.47 ± 9.07	1.06×10^{-3}
avg.				

Table 4: Average utility-discounted accuracies (%) and time to predict in seconds.

First, we can see that including some cautiousness can increase our accuracies on most data sets, by picking the right values of c. This increase is sometimes noticeable, for example in the vehicle (i), wine-quality (l), wall-following (o) and vowel data sets (j). All of this, keeping a time execution reasonable in view of the problems to be solved (e.g. a non-convex, NP-hard problem), and without an optimized implementation. As expected, assuming independence between the features (i.e., diagonal covariance matrices) significantly reduces the computational time, making it negligible, but overall reduces performances, as the assumptions are often violated in a stronger way.

In order to highlight the major role of cautiousness of an imprecise classifier model, we show in Figure 6(c) and 6(b) how, in the IRIS data set, our IQDA and ILDA models create different areas of decision boundaries (not to be confused with rejection area), where each area has a different combinations of subset of labels $\widehat{Y} \subseteq \mathscr{K}$, in contrast to precise classifier model (LDA), in Figure 6(a), where it creates one area per label. We can clearly see that the two classifiers behave quite differently. In particular, ILDA will induces

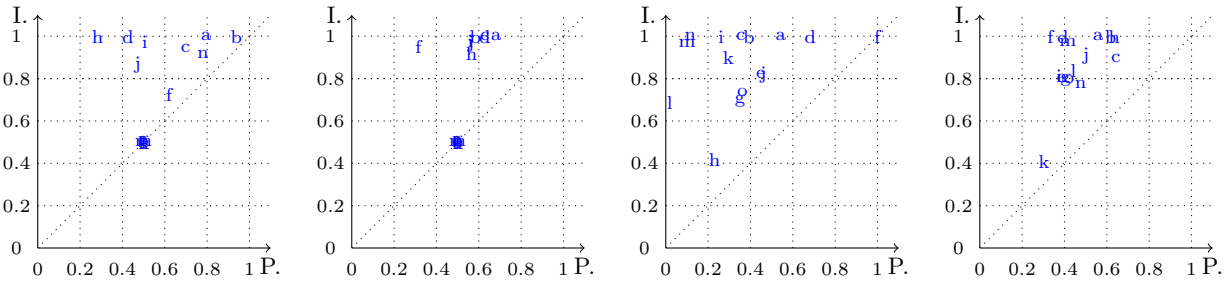


(a) Precise boundary decision (b) ILDA zone decision (c) IQDA zone decision

Figure 6: Figure 6(a) shows how a precise model divides the instance space in three single different zones by label (i.e $\{a\}, \{b\}, \{c\}$), the Figure 6(b) shows how an ILDA model divides the instance space in different zones as much as different combinations of a subset of labels (i.e $\{a\}, \{b\}, \{c\}, \{a, b\}, \{b, c\}$, and so on), and the Figure 6(c) shows how IQDA model can also divide in different zones with smooth curves instead.

regions delimited by piece-wise linear functions, while IQDA will induces regions delimited by piece-wise quadratic functions.

Also, in Figure 8, we show the evolution of utility-discounted accuracy (i.e. u_{65} and u_{80} of vowel dataset), with a standard deviation calculated by a 10-fold cross-validation on the training dataset, according to the imprecision of estimators μ. As expected we notice that when c reaches a too high value, the overall model performances decrease, as it becomes too imprecise with respect to our attitude towards cautiousness (modelled through utility (41)). The rest of experiments are in Appendix A.

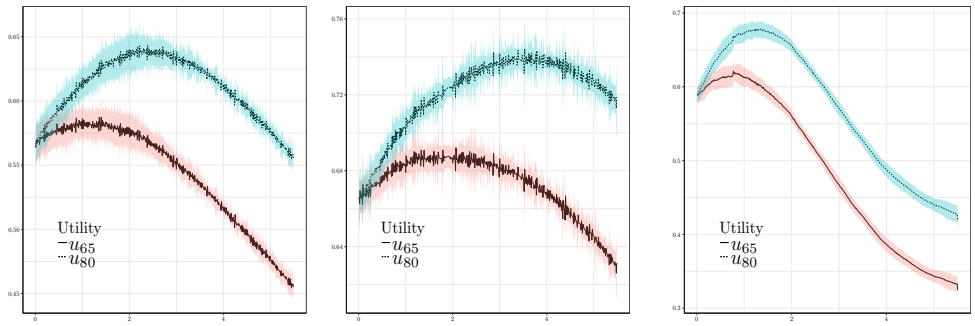


(a) ILDA & LDA (b) IQDA & QDA (c) IEDA & EDA (d) INDA & NDA

Figure 7: Correctness of the different methods in the case of abstention versus accuracy of their precise counterparts, only on those instances for which an indeterminate prediction was given. Graphs are given for the u_{80} accuracies.

An imprecise classifier should abstain (i.e. by providing a set of plausible choices) on those hard instances, that is the instances where the *precise* classifier makes a unusual high amount of mistakes. In Figure 7, we verify that our imprecise classifiers follow this desirable behaviour on most data sets, for the u_{80} measures (conclusions for the u_{60} are similar, but not displayed to gain some space). Figure 7(a) displays the percentage of time the true label is in the prediction of ILDA, given that the prediction was imprecise, versus the accuracy of LDA on those same instances. The same graphs for the QLDA, IEDA and INDA methods are given by Figure 7(b), Figure 7(c) and Figure 7(d), respectively. We notice that on those hard instances where precise classifiers are wrong, our imprecise classifiers successfully overcome them, getting the ground-truth value into partial predictions (most often $> 80\%$). A typical and quite remarkable example of this is the dermatology data set (h) for the linear case, where the accuracy on the imprecisely classified instances drop to 30% for the precise classifier (to be compared to an average of 96% on all instances), while the imprecise classifier always include the true class. Moreover, the fact that u_{80} is higher indicates that the overall amount of imprecision remains acceptable. Our approach therefore seems to be able to well robustify

the very simple, linear decision frontiers of the ILDA models.



(a) IEDA model (b) INDA model (c) IQDA model (d) ILDA model

Figure 8: Figures shows performance evolution with a standard deviation region of three principales methods, (1)Figure 8(d) for ILDA model, (2) Figure 8(c) for IQDA model, and (3) Figure 8(b) for INDA model, w.r.t. utility-discount accuracy u_{65}, u_{80} and c tuning parameter on vowel dataset

Before considering some generalisation of the presented methods, we would also like to mention that the imprecise probabilistic approach will in general induces decision frontiers that are different from classical rejection rule. Figure 5(b) illustrates this well: rejection regions in a binary setting are most often equivalent to require to predict $\{a, b\}$ whenever $\hat{P}(\{a\}|x) \in [0.5 - \epsilon, 0.5 + \epsilon]$ for some ϵ. This means that in the case of LDA, the rejection regions will be delimited by two parallel lines, corresponding to the iso-density points x for which $\hat{P}(\{a\}|x) = 0.5 - \epsilon$ and $\hat{P}(\{a\}|x) = 0.5 + \epsilon$. In contrast, we can clearly see in Figure 5(b) that the boundaries are not linear, but piece-wise linear.

6. Imprecise prior marginal and generic loss functions

In this section, we will discuss about two new variants of IGDA model: (1) relaxing the Assumption 2, i.e. $\mathbb{P}_Y := \widehat{\pi}$, with the purpose of putting a set of probability distributions $\mathscr{P}_Y$ instead, and (2) dealing with generic loss function instead classical $\mathscr{L}_{0/1}$ loss function. We will evaluate the impact of this two new variants in our IGDA model in terms of added computational complexity.

6.1. Imprecise prior marginal

The first extension we will consider is to make imprecise the marginal distribution, considering a set $\mathscr{P}_Y$ rather than a precise distribtuion, in the same vein as we have made the conditional distribution $\mathscr{P}_{X|Y}$ imprecise. For the time being, we will still work with the $\mathscr{L}_{0/1}$ loss function. Since the conditionals are still independent of each other, solving the maximality criterion amounts to solve Equation (25), that we recall here

$$\inf_{\mathbb{P}_Y \in \mathscr{P}_Y} \underline{P}(X = \boldsymbol{x}|Y = m_a)P(Y = m_a) - \overline{P}(X = \boldsymbol{x}|Y = m_b)P(Y = m_b) \tag{42}$$

with $m_a \succ_M m_b$ if this is positive. This equation can be solved easily, as it is a linear form in $P(Y = m_a), P(Y = m_b)$, meaning that we can either use linear programming over the constraints induced by $\mathscr{P}_Y$, or find the extreme point (e.g., by enumeration) of $\mathscr{P}_Y$ for which the solution is obtained. Recall t

The problem then amounts to estimate $\mathscr{P}_Y$. A quite popular choice to do so is to use an Imprecise Dirichlet Model (IDM) [38, 39]. However, as Benavoli *et al.* has already mentioned in [18, §4.2], the set of prior distributions of IDM does not correctly satisfy (P1) Prior-invariance property and permutation invariance of near-ignorance model. So, to remain consistent with our previous estimates, we explore another solution proposed by Benavoli *et al.*.

Let Y be a discrete random variable on a finite space of labels $\mathscr{K}$ with probability distribution $\mathbb{P}_Y$ and let the parameters $\pi_{m_k}, \forall m_k \in \mathcal{K}$ be the unknown non-negative chances, i.e. $P(Y = m_k)$. The Corollary 4.10 in [18] propose adding some contraints in the space $\mathbb{L}$ in order not to favour some chances π_{m_k} over others. They then consider the following set of prior distributions:

$$\mathscr{P}_\pi = \left\{ \pi_{m_1}^{\ell_1 - 1} \pi_{m_2}^{\ell_2 - 1} \dots \pi_{m_K}^{-\sum_{i=1}^{d-1} \ell_i - 1}, \|\ell\|_1 \leq 2c, \sum_{i=1}^{d-1} \ell_i \in [-c, c] \right\}. \tag{43}$$

It is also shown [18, Eq. 24] that, after combining this set with the likelihood, the lower and upper expectations of the chances of observing a given subset A of categories result in

$$\overline{\mathbb{E}}\left[\sum\nolimits_{m_k \in A} \pi_{m_k} \middle| n, \hat{\mathbf{y}}_{\mathbf{n}} \right] = \min\left(1, \frac{1}{n}\left[\sum\nolimits_{m_k \in A} n_k + c \right]\right) := \underline{P}_Y(A), \tag{44}$$

$$\underline{\mathbb{E}}\left[\sum\nolimits_{m_k \in A} \pi_{m_k} \middle| n, \hat{\mathbf{y}}_{\mathbf{n}} \right] = \max\left(0, \frac{1}{n}\left[\sum\nolimits_{m_k \in A} n_k - c \right]\right) := \overline{P}_Y(A), \tag{45}$$

where n is the total number of observations in the data set, i.e $n = |\mathscr{D}| = N$. We will then consider the probability set

$$\mathscr{P}_Y = \left\{ P \middle| \underline{P}_Y(A) \leq P(A) \leq \overline{P}_Y(A), \forall A \subseteq \mathcal{K} \right\} \tag{46}$$

Such a model, which corresponds to take a neighbourhood around the empirical distribution using the total variation distance (i.e., L_∞ norm) has been recently investigated by Miranda *et al.* [40], showing for instance that it induced a 2-monotone lower probability, but was not a specific case of probability intervals, in contrast with the IDM model. Using this fact, we know that the result of Equation (25) will be obtained by the Choquet integral, which results in this particular case in

$$\begin{aligned}
&\inf_{\mathbb{P}_Y \in \mathscr{P}_Y} \underline{P}(X = \boldsymbol{x} | Y = m_a) P(Y = m_a) - \overline{P}(X = \boldsymbol{x} | Y = m_b) P(Y = m_b) = \\
&\underline{P}(X = \boldsymbol{x} | Y = m_a) \underline{P}_Y(\{m_a\}) - \overline{P}(X = \boldsymbol{x} | Y = m_b) \overline{P}_Y(Y = m_b) = \\
&\underline{P}(X = \boldsymbol{x} | Y = m_a) \max\left(0, \frac{n_a - c}{n}\right) - \overline{P}(X = \boldsymbol{x} | Y = m_b) \min\left(1, \frac{n_b + c}{n}\right)
\end{aligned}$$

In particular, this shows that there would be no differences if we considered only the projections of $\mathscr{P}_Y$ over its singletons, which amounts to consider the bigger set

$$\mathscr{P}'_Y = \left\{ P(Y = m_k) = \pi_{m_k} \middle| \pi_{m_k} \in \left[\max\left(0, \frac{n_k - c}{n}\right), \min\left(1, \frac{n_k + c}{n}\right) \right] \forall m_k \in \mathcal{K} \right\}. \tag{47}$$

6.2. Generic loss function

The *zero-one* loss function is the default loss function used in classification problems (where we consider that penalty of being wrong is the same for every kind of error). However, in many practical problems different errors will have different impacts, and this is especially true for sensitive applications in which imprecise probabilistic approaches could be useful.

Equation Equation (9) can be written

$$m_a \succ_M m_b \iff \inf_{\substack{\mathbb{P}_{X|m_*} \in \mathscr{P}_{X|m_*} \\ \mathbb{P}_Y \in \mathscr{P}_Y}} \sum_{m_k \in \mathcal{K}} \left(\mathcal{L}(m_k, m_b) - \mathcal{L}(m_k, m_a) \right) P(Y = m_k | X = \boldsymbol{x}) > 0, \tag{48}$$

which, if we denote by $c_{m_k}^{b-a} := \mathcal{L}(m_k, m_b) - \mathcal{L}(m_k, m_a)$, gives

$$\iff \inf_{\substack{\mathbb{P}_{X|m_*} \in \mathscr{P}_{X|m_*} \\ \mathbb{P}_Y \in \mathscr{P}_Y}} \sum_{m_k \in \mathcal{K}} c_{m_k}^{b-a} P(X = \boldsymbol{x}|Y = m_k) P(Y = m_k) > 0 \tag{49}$$

$$\iff \inf_{\mathbb{P}_Y \in \mathscr{P}_Y} \sum_{\{k|c_{m_k}^{b-a}>0\}} c_{m_k}^{b-a} \underline{P}(X = \boldsymbol{x}|Y = m_k) P(Y = m_k) + \sum_{\{k|c_{m_k}^{b-a}\leq 0\}} c_{m_k}^{b-a} \overline{P}(X = \boldsymbol{x}|Y = m_k) P(Y = m_k) > 0 \tag{50}$$

that uses the fact that the conditional probabilities $P(X = \boldsymbol{x}|Y = m_k)$ are all independent. As Equation (50) remains a linear form of the probabilities $P(Y = m_k)$, it can be solved as previously, i.e., thourgh the use of linear programming or the identification of the extreme point for which the bound is reached.

If we now consider the credal set given by Equation (46) and induced by the constraints (44)-(45), we still have that this induces a 2-monotone lower probability, meaning that we can estimate (50) by using the Choquet integral.

All these remarks show that making the marginal probabilities imprecise or considering generic loss functions does not make the model more complex to use, as the computational complexity is not increased by much, especially when $\mathscr{P}_Y$ has mathematical properties making computations easier (which is luckily the case for most IP models over multinomial distributions).

7. Conclusion

In this paper, we have generalized classical Gaussian discriminant models to the imprecise setting, mainly by allowing the estimated means of the conditional Gaussian distributions to become imprecise. This was achieved by a robust Bayesian procedure using sets of prior satisfying near-ignorance properties.

We have explored the computational issues associated to the predictions of such models, essentially showing that considering general covariance matrices ended up in practically manageable but theoretically difficult to solve problems, while considering diagonal covariance matrices essentially made the problem much easier to solve.

Experiments on various data sets shows that the method is providing quite satisfactory results, in the sense that the induced imprecision in the predictions is reasonable and mostly concerns instances that were wrongly classified by the precise methods. We have also discussed some possible extensions of our approaches, showing that such extensions would not add a prohibitive computational cost.

A natural next step would be to also make the covariance matrix estimate imprecise, possibly leaving the mean estimate precise in a first step. Computationally, this would be attractive, as the objective functions are a mainly linear of the covariances if the mean is left imprecise. The main problem would then be to derive a principled approach (i.e., using near-ignorance prior) that would deliver an easy-to-deal convex set of inverse covariance matrices.

Acknowledgements

This work was carried out in the framework of the Labex MS2T, funded by the French Government, through the National Agency for Research (Reference ANR-11-IDEX-0004-02).

Appendix A. Experimental results

Complementary experimental results are shown in the Figure A.9 and Figure A.10

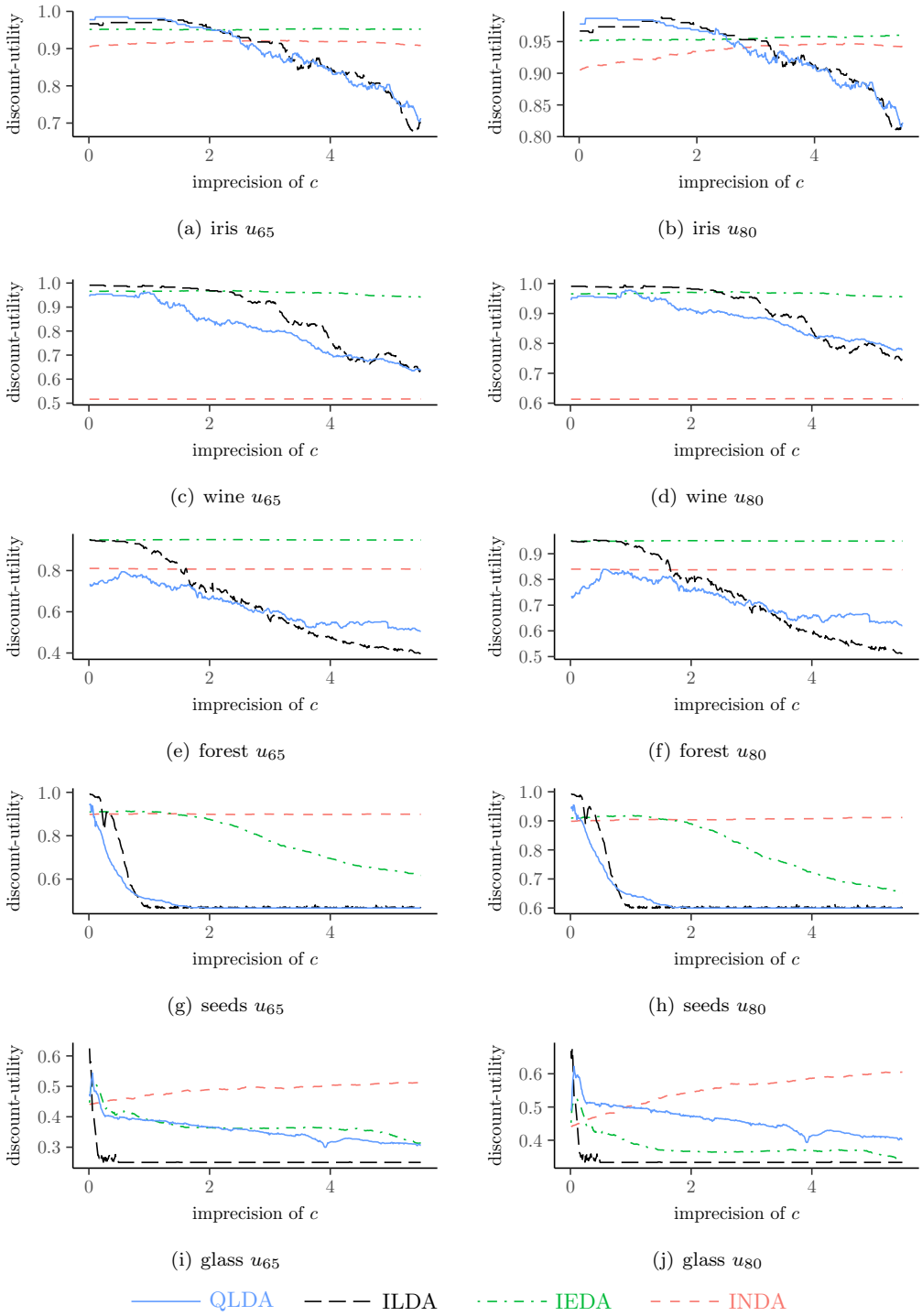


Figure A.9: Experiments for IGDA model (left:utility-discount u_{65}, right:utility-discount u_{80})

References

References

[1] R. Kitchin, T. P. Lauriault, Small data in the era of big data, GeoJournal 80 (4) (2015) 463–475.

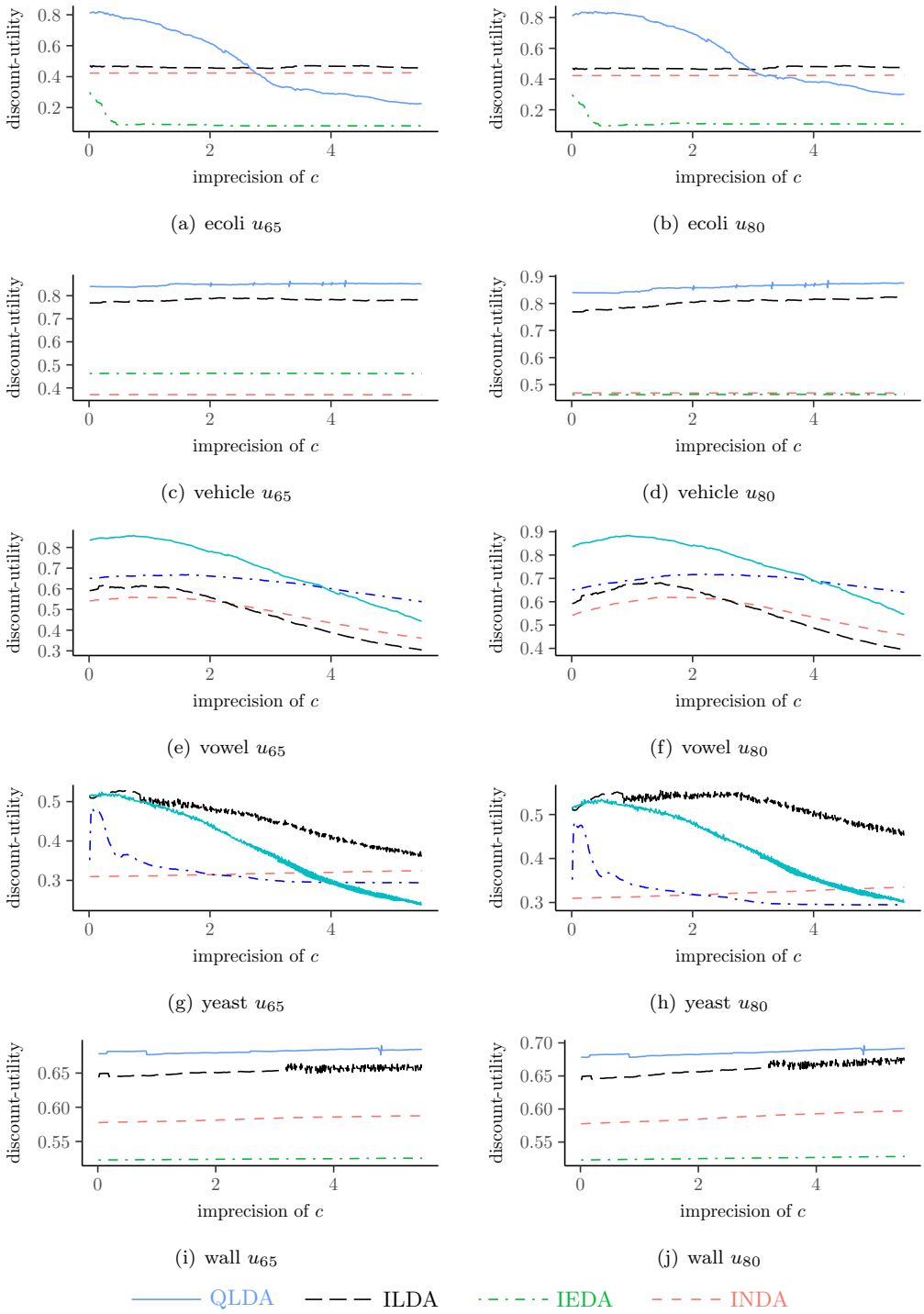


Figure A.10: Experiments for IGDA model (left:utility-discount u_{65}, right:utility-discount u_{80})

[2] L. A. Dalton, M. R. Yousefi, On optimal bayesian classification and risk estimation under multiple classes, EURASIP Journal on Bioinformatics and Systems Biology 2015 (1) (2015) 8.

[3] S. Roeser, R. Hillerbrand, P. Sandin, M. Peterson, Essentials of risk theory, Springer Science & Business Media, 2012.

[4] R. Senge, S. Bösner, K. Dembczyński, J. Haasenritter, O. Hirsch, N. Donner-Banzhoff, E. Hüllermeier, Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty, Information Sciences 255 (2014) 16–29.
[5] U. M. Braga-Neto, E. R. Dougherty, Is cross-validation valid for small-sample microarray classification?, Bioinformatics 20 (3) (2004) 374–380.
[6] J. Friedman, T. Hastie, R. Tibshirani, The elements of statistical learning, Springer New York Inc., 2001.
[7] T. Augustin, F. P. Coolen, G. de Cooman, M. C. Troffaes, Introduction to imprecise probabilities, John Wiley & Sons, 2014.
[8] R. Herbei, M. H. Wegkamp, Classification with reject option, Canadian Journal of Statistics 34 (4) (2006) 709–721.
[9] W. Cheng, E. Hüllermeier, W. Waegeman, V. Welker, Label ranking with partial abstention based on thresholded probabilistic models, in: Advances in neural information processing systems, 2012, pp. 2501–2509.
[10] T. M. Ha, The optimum class-selective rejection rule, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (6) (1997) 608–615.
[11] G. Shafer, V. Vovk, A tutorial on conformal prediction, Journal of Machine Learning Research 9 (Mar) (2008) 371–421.
[12] M. E. Cattaneo, Statistical decisions based directly on the likelihood function, Ph.D. thesis, ETH Zurich (2007).
[13] M. E. Cattaneo, Fuzzy probabilities based on the likelihood function, in: Soft Methods for Handling Variability and Imprecision, Springer, 2008, pp. 43–50.
[14] P. Walley, Statistical reasoning with imprecise Probabilities, Chapman and Hall, 1991.
[15] G. Walter, Generalized bayesian inference under prior-data conflict, Ph.D. thesis, lmu (2013).
[16] E. Quaeghebeur, G. De Cooman, Imprecise probability models for inference in exponential families, in: 4th International Symposium on Imprecise Probabilities and Their Applications, International Society for Imprecise Probability: Theories and Applications (SIPTA), 2005, pp. 287–296.
[17] J. M. Bernardo, A. F. Smith, Bayesian Theory, John Wiley & Sons Ltd., 2000.
[18] A. Benavoli, M. Zaffalon, Prior near ignorance for inferences in the k-parameter exponential family, Statistics 49 (5) (2014) 1104–1140.
[19] I. Levi, The enterprise of knowledge: An essay on knowledge, credal probability, and chance, MIT press, 1983.
[20] J. O. Berger, Statistical decision theory and Bayesian analysis; 2nd ed., Springer Series in Statistics, Springer, New York, 1985.
[21] M. C. Troffaes, Decision making under uncertainty using imprecise probabilities, International Journal of Approximate Reasoning 45 (1) (2007) 17–29.
[22] A. P. Dempster, A generalization of bayesian inference, Journal of the Royal Statistical Society: Series B (Methodological) 30 (2) (1968) 205–232.
[23] S. J. Taylor, Introduction to measure and integration, CUP Archive, 1973.
[24] M. Zaffalon, The naive credal classifier, Journal of statistical planning and inference 105 (1) (2002) 5–21.
[25] G. Yang, S. Destercke, M.-H. Masson, Cautious classification with nested dichotomies and imprecise probabilities, Soft Computing 21 (24) (2017) 7447–7462.
[26] A. Benavoli, B. Ristic, Classification with imprecise likelihoods: A comparison of tbm, random set and imprecise probability approach, in: Proceedings of the 14th International Conference on Information Fusion, IEEE, 2011, pp. 1–8.
[27] V. R. Marco, D. M. Young, D. W. Turner, The euclidean distance classifier: an alternative to the linear discriminant function, Communications in Statistics-Simulation and Computation 16 (2) (1987) 485–505.
[28] C. Robert, Le choix bayésien: Principes et pratique, Springer Paris, 2005.
[29] P. L. De Angelis, P. M. Pardalos, G. Toraldo, Quadratic programming with box constraints, in: Developments in global optimization, Springer US, 1997, pp. 73–93.
[30] C. Johnson, Positive definite matrices, The American Mathematical Monthly 77 (3) (1970) 259–264.
[31] M. S. Andersen, J. Dahl, L. Vandenberghe, Cvxopt: A python package for convex optimization, version 1.2.2, Available at cvxopt.org.
[32] P. M. Pardalos, S. A. Vavasis, Quadratic programming with one negative eigenvalue is np-hard, Journal of Global Optimization 1 (1) (1991) 15–22.
[33] S. Burer, D. Vandenbussche, Globally solving box-constrained nonconvex quadratic programs with semidefinite-based finite branch-and-bound, Computational Optimization and Applications 43 (2) (2009) 181–195.
[34] W. Xia, J. Vera, L. F. Zuluaga, Globally solving non-convex quadratic programs via linear integer programming techniques, arXiv preprint arXiv:1511.02423.
[35] A. Frank, A. Asuncion, UCI machine learning repository (2010).
URL http://archive.ics.uci.edu/ml
[36] J. H. Friedman, Regularized discriminant analysis, Journal of the American statistical association 84 (405) (1989) 165–175.
[37] M. Zaffalon, G. Corani, D. Mauá, Evaluating credal classifiers by utility-discounted predictive accuracy, International Journal of Approximate Reasoning 53 (8) (2012) 1282–1301.
[38] P. Walley, Inferences from multinomial data: learning about a bag of marbles, Journal of the Royal Statistical Society. Series B (Methodological) (1996) 3–57.
[39] J.-M. Bernard, An introduction to the imprecise dirichlet model for multinomial data, International Journal of Approximate Reasoning 39 (2-3) (2005) 123–150.
[40] E. Miranda, I. Montes, S. Destercke, A unifying frame for neighbourhood and distortion models, in: International Symposium on Imprecise Probability: Theories and Applications (ISIPTA), 2019.