Imprecise Gaussian Discriminant Classification

Yonatan-Carlos Carranza-Alarcon YONATAN-CARLOS.CARRANZA-ALARCON@HDS.UTC.FR
Sébastien Destercke SEBASTIEN.DESTERCKE@HDS.UTC.FR
Sorbonne Universités, Université Technologique de Compiègne, CNRS, UMR 7253 - Heudiasyc, 57 Avenue de Landshut, Compiègne, France

Abstract

Gaussian discriminant analysis is a popular classification model, that in the precise case can produce unreliable predictions in case of high uncertainty. While imprecise probability theory offer a nice theoretical framework to solve this issue, 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 produces reasonably cautious predictions, in the sense that the number of set-valued predictions is not too high, and that those predictions correspond to hard-to-classify instances, that is instances for which the precise classifier accuracy drops.

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 $\mathcal{K}$ of such labels) a new instance $\boldsymbol{x} \in \mathcal{X}$ belongs. The reliability of this *precise* prediction may depend heavily on prior beliefs (e.g., assumptions made by data analysts) and the nature of training data set (e.g., in small amounts [15, 8] and/or with high degree of *uncertainty*). A well-known generative model used to perform the classification task is the Gaussian discriminant analysis (GDA) [12, §4.3].

Let $\mathcal{X} \times \mathcal{K}$ be the space of observations and possible labels, with $X \in \mathcal{X} = \mathbb{R}^p$ a multivariate random variable and $Y \in \mathcal{K} = \{m_1, \ldots, m_K\}$ the set of labels. 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

$$\mathbb{P}_{Y=m_k|X} = \frac{\mathbb{P}_{X|Y=m_k}\mathbb{P}_{Y=m_k}}{\sum_{m_l \in \mathcal{K}} \mathbb{P}_{X|Y=m_l}\mathbb{P}_{Y=m_l}}. \quad (1)$$

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 $\mathbb{P}_{Y=m_k|X}$, such as assuming that they are Gaussian probability distributions (g.p.d), in order to find a plausible estimate (see Section 3.1). However, such precise estimates usually have trouble differentiating various kinds of uncertainties [19], 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 [6]). 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 illustration in Figures 1(*a*) and 1(*b*)).

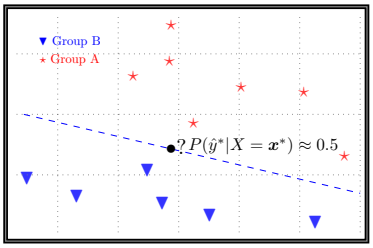
While Bayesian methods may mitigate the impact of limited information by using prior distributions, such prior distributions arguably contain themselves a lot of information. In order to properly model the absence of prior beliefs, Walley [22, §4.6.9] proposed to use the generalized notion of *near-ignorance prior*, which must respect certain properties [3, §2] we would expect from an uncertainty model of ignorance, while allowing one to learn from data.

Applied to classification problems, such approaches result in Imprecise classifiers that do not aim to do "better" than their precise counterparts, nor to implement a rejection option (i.e., not classifying at all) in case of ambiguity [13], 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.

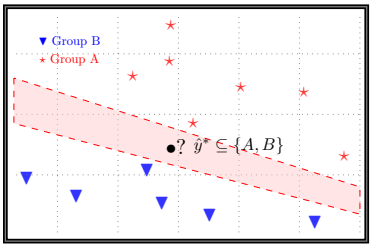
In this paper, we propose an extension of GDA model to a new (cautious[1]) *imprecise* classification model, named Imprecise GDA (IGDA), based on a robust Bayesian and near-ignorance approach proposed in [3] and concentrating on the imprecise estimation of the *means* μ of a set of g.p.d. In other words, IGDA aims to better describe the lack of evidence derived from limited information in the data, resulting in a set of conditional distributions (or credal set [16]) for every m_k, noted $\mathscr{P}_{X|Y=m_k}$, instead of a single c.p.d $\mathbb{P}_{X|Y=m_k}$.

In the following, we will first provide in Section 2 the basics of *precise* classification learning from the point of view of statistical decision theory [12, §2] and preference ordering of utility theory [4, §2.2], as well as the impre-

1. Cautious and imprecise are here used interchangeably.



(*a*) Precise decision-making



(*b*) Cautious decisions-making

Figure 1: *Cautious vs precise decision-making.* We can remark in the figure (a) that precise model can produce many mistakes for hard to predict unlabeled instances, in contrast to cautious model (b) where it recognizes such instances and makes a cautious decision.

cise probabilistic extension of *precise* decision called the *maximality* criterion [21]. In Section 3, we describe the estimation of the model, both in the *precise* and *imprecise* settings. Integrating this *imprecise* estimation with *maximality* criterion, we will present two variants of our IGDA model in Section 4.

Finally, in Section 5, we perform some experiments with different datasets and show a comparative table of results with precise discriminant analysis model.

2. Preliminaries

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 $\mathscr{X} \times \mathscr{K}$. In statistical decision theory, the aim of classification is to learn a model $\varphi : \mathscr{X} \to \mathscr{K}$ that minimizes the risk of misclassification with respect to a loss function $\mathscr{L}(\cdot,\cdot) : \mathscr{K} \times \mathscr{K} \to \mathbb{R}$. Such an optimal predictive model can be defined as

$$\hat{\varphi} := \arg\min_{y \in \mathscr{K}} \mathbb{E}_{Y|X} \left[\mathscr{L}(y, \varphi(X)) \right] \tag{2}$$

When $\mathscr{L}$ is the classical *zero-one* loss function, where $\mathscr{L}_{0/1}(y, \hat{y})$ is equal to 1 if $y \neq \hat{y}$ (with $\hat{y} := \hat{\varphi}(\boldsymbol{x})$ obtained by Equation (2)) and 0 otherwise, Equation (2) comes down to choose as $\widehat{\varphi}(\cdot)$ the class maximizing the conditional probability of Y given the unlabeled instance $\boldsymbol{x} \notin \mathscr{D}$:

$$\widehat{\varphi}(\boldsymbol{x}) = \text{argmax}_{m_k \in \mathscr{K}} \, P(Y = m_k | X = \boldsymbol{x}). \tag{3}$$

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 1 (Precise ordering [4, pp. 47]) *Given a general loss function $\mathscr{L}(\cdot,\cdot)$, a conditional probability distribution $\mathbb{P}_{Y|\boldsymbol{x}}$ and a new unlabeled instance $\boldsymbol{x}$, m_a is preferred to m_b , denoted by $m_a \succ m_b$, if and only if:*

$$\mathbb{E}_{\mathbb{P}_{Y|\boldsymbol{x}}} [\mathscr{L}(\cdot, m_a)] < \mathbb{E}_{\mathbb{P}_{Y|\boldsymbol{x}}} [\mathscr{L}(\cdot, m_b)] \tag{4}$$

Definition 1 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 of m_a. In the particular case where we use the loss function $\mathscr{L}_{0/1}$, it is easy to prove that:

$$m_a \succ m_b \iff P(Y = m_a | X = \boldsymbol{x}) > P(Y = m_b | X = \boldsymbol{x}) \tag{5}$$

where $P(Y = m_a | X = \boldsymbol{x})$ is the unknown conditional probability of label m_a given a new unlabeled instance $\boldsymbol{x}$. Therefore, given a set of labels $\mathscr{K}$, we can then establish a complete preorder making pairwise comparisons (see figure 2), and then, picking out the label which is maximal in that pre-order as a final decision.

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

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

the complete preorder over labels w.r.t. estimated probabilities is $m_c \succ m_a \succ m_b$ where m_c is the maximal predicted label dominating all others (Figure 2).

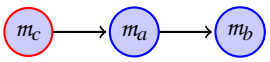


Figure 2: Graph of complete preorder on labels

Hence, whether we consider Equation (3) or (5), the main task in *precise* probabilistic approaches is to estimate the *single* conditional distribution $\mathbb{P}_{Y|X}$ which usually models the *uncertainty* in data. Yet, several authors [22, 10] have argued that a single distribution cannot always faithfully represent all uncertainties, and in particular lack of information, and recommend using sets of probabilities (or equivalent models) to represent it. This is the approach we

consider here, in order to derive a partial prediction in the form of set of labels $\widehat{Y} \subseteq \mathscr{K}$ instead of a *precise* label $\widehat{y}$ when we are unsure about the optimal decision.

2.2. Imprecise Probabilities

Imprecise probabilities consist in representing our uncertainty by a convex set $\mathscr{P}_X$ of probabilities [22, 1], defined over a space $\mathscr{X}$, rather than by a precise probability measure $\mathbb{P}_X$ [20].

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:

$$\underline{P}_X(A) = \inf_{P \in \mathscr{P}_X} P(A) \quad \text{and} \quad \overline{P}_X(A) = \sup_{P \in \mathscr{P}_X} P(A) \tag{6}$$

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 distribution instead of a *single* distribution $\mathbb{P}_X$. That is why we rely in our case on an efficient inference model, the generalized Bayesian inference method proposed by Benavoli and Zaffalon [3] (or robust Bayesian inference) for exponential families, which we will present in Section 3.2. For theoretical developments of the next subsection, we will assume that we already know the set $\mathscr{P}_{Y|X}$ of conditional distributions.

2.3. Decision Making under Imprecise Probabilities

In the context of imprecise probabilities, we can find different methods extending the decision criterion given in Definition 1 (for more details, see Troffaes [21]). For classifying a new instance $\boldsymbol{x}$, we will use the *maximality criterion* [1, §8.6] that benefits from strong theoretical justifications [22, §3.9.5] and often remains applicable in practice [25, 24, 2]. It extends Equation (4) and is defined as follows:

Definition 2 (Partial Ordering by Maximality Criterion [21, §3.2]) *Let $\mathscr{L}(\cdot,\cdot)$ be a general loss function, $\boldsymbol{x}$ an observed instance and $\mathscr{P}_{Y|\boldsymbol{x}}$ a set of conditional probability distributions. m_a is preferred to m_b according to the maximility criterion if the cost of exchanging m_a with m_b has a positive lower expectation:*

$$m_a \succ_M m_b \iff \inf_{\mathbb{P}_{Y|\boldsymbol{x}} \in \mathscr{P}_{Y|\boldsymbol{x}}} \mathbb{E}_{\mathbb{P}_{Y|\boldsymbol{x}}}\left[\mathscr{L}(\cdot, m_b) - \mathscr{L}(\cdot, m_a)\right] > 0 \tag{7}$$

if $\mathscr{L}(\cdot,\cdot)$ is $0/1$ loss function, $m_a \succ_M m_b$ if and only if:

$$\inf_{\mathbb{P}_{Y|\boldsymbol{x}} \in \mathscr{P}_{Y|\boldsymbol{x}}} P(Y = m_a|\boldsymbol{x}) - P(Y = m_b|\boldsymbol{x}) > 0 \tag{8}$$

Equation (8) amounts to asking that Equation (4) 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 $N \to \infty$, cautious and precise models will usually coincide. The prediction $\widehat{Y}_M$ resulting from $\succ_M$ is defined as:

$$\widehat{Y}_M = \left\{ m_a \in \mathscr{K} \;\middle|\; \nexists m_b \in \mathscr{K} : m_b \succ_M m_a \right\} \tag{9}$$

Example 2 *If our label space is $\mathscr{K} = \{m_a, m_b, m_c\}$, a possible partial ordering could be the following:*

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

where $\widehat{Y}_M = \{m_a, m_c\}$ is the predicted set obtained from the set $\mathscr{B}$ of comparisons derived by the criterion of maximality (Figure 3).

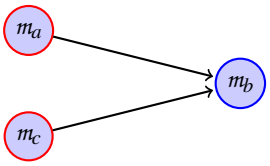


Figure 3: Graph of partial order of set $\mathscr{B}$.

3. Gaussian Discriminant Analysis Model

A classical way, already mentioned in the introduction, to estimate the distribution $\mathbb{P}_{Y|X}$ is by using Bayes' theorem and by assuming a specific form for $\mathbb{P}_{X|Y=m_k}$, in our case a normality assumption. Making use of Equality (1), we will discuss first the precise and then the retained imprecise approach, respectively in Sections 3.1 and 3.2.

3.1. Statistical Inference with Precise Probabilities

GDA focuses on a *parametric* estimation assuming that $\mathbb{P}_{X|Y=m_k}$ follows a multivariate Gaussian distribution $\mathscr{N}(\mu_{m_k}, \Sigma_{m_k})$ with mean μ_{m_k} and covariance matrix Σ_{m_k}, i.e.:

$$\mathscr{G}_{m_k} := \mathbb{P}_{X|Y=m_k} \sim \mathscr{N}(\mu_{m_k}, \Sigma_{m_k}) \tag{10}$$

where their probability density function is written:

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

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

$$\underset{m_k \in \mathscr{K}}{\arg\max} \; \log \pi_{m_k} - \log|\Sigma_{m_k}|^{\frac{1}{2}} - \frac{1}{2}(\boldsymbol{x}^T - \mu_{m_k})^T \Sigma_{m_k}^{-1}(\boldsymbol{x}^T - \mu_{m_k}) \tag{11}$$

where $\Theta = \{\theta_{m_k} | \theta_{m_k} = (\pi_{m_k}, \Sigma_{m_k}, \mu_{m_k}), \forall m_k \in \mathscr{K}\}$ is our parametric space.

In frequentist inference, usual estimation of parameters of (11) is obtained by MLE using a subset $\mathscr{D}_{m_k} = \{(x_{i,k}, y_{i,k}{=}m_k) | i = 1, \ldots, n_{m_k}\} \subseteq \mathscr{D}$ of observations of training data. We have $\hat{\pi}_{m_k} = n_{m_k}/N$ (frequency of m_k) and

$\hat{\mu}_{m_k} = \bar{\boldsymbol{x}}_{m_k}$ (sample mean of $\mathscr{D}_{m_k}$). Depending on whether we assume hetero or homoscedasticity, we have 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 $\mathscr{D}$). They are respectively known as Quadratic and Linear discriminant model [12, §4.3].

Those estimates do not account for the quantity of data they are based on, which may be low to start with, and will 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.* [3].

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 approach. Before describing our *imprecise* estimates, we make three general assumptions for our *imprecise* Gaussian discriminant model:

1. Hypothesis of 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 $\Sigma_{m_k} := \widehat{\Sigma}_{m_k} = \widehat{S}_{m_k}$ or $\widehat{S}$ (resp. homo- or hetero- scedastic case)

Relaxing these assumptions (in particular 3) is the matter of future works.

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 about the 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 $\prod_i^{n_{m_k}} P_{x_{i,m_k}|y_{i,m_k},\theta_{m_k}}$ 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 of ***mean parameters*** (i.e. $\theta_{m_k} = \mu_{m_k}$), assuming a (precise) estimation of $\hat{\Sigma}_{m_k}$. Thus, the posterior on the mean is such that

$$P(\mu_{m_k} \mid \mathscr{D}_{m_k}) \propto \prod_i^{n_{m_k}} P(X = \boldsymbol{x}_{i,m_k} \mid \mu_{m_k}, \boldsymbol{y}_{i,m_k}) P(\mu_{m_k}). \quad (12)$$

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 $\mathscr{D}_{m_k}$ of label m_k.

To make imprecise estimations in the form of convex set, we will use a set of prior distribution $\mathscr{P}_\mu$, leading to a set of posterior distributions $\mathscr{P}_{\mu|\cdot}$. Besides, as we do not have any prior belief about the unknown parameters μ, we will use a set of prior distributions that represent this absence of prior knowledge, while still allowing for learning, known as "near-ignorance" priors.

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*. *Benavoli et al* in [3] propose a new near-ignorance model, about a multivariate random variable based on a set of distribution $\mathscr{M}$, which aims to reconcile two approaches, namely, re-parametrization invariance and *Walley's* near-ignorance prior. For that, *Benavoli 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 [3, §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 not 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 the limit.

Benavoli et al in [3] provide a set of conjugate priors $\mathscr{M}$ for *regular multivariate exponential families* [18, §3.3.4] ($\mathscr{FE}xp$) that satisfy the last four properties under quite weak assumptions. Borrowing from [3], we can define this set of prior distributions $\mathscr{M}$ as follows:

Definition 3 (Prior near-ignorance for k-parameter exponential families [3, §4, eq. 16]) *Let $\mathbb{L}$ be a bounded closed convex subset of $\mathbb{R}^k$ strictly including the origin ([3, lem. 4.5]).*

$$\mathbb{L} = \left\{ \ell \in \mathbb{R}^k : \ell_i \in [-c_i, c_i], c_i > 0, i = \{1, \ldots, d\} \right\} \quad (13)$$

Let $W \in \mathscr{W} = \mathbb{R}^k$ be a random variable with probability density function, if and only if for all $\ell_i \neq 0$:

$$p(w) = exp(\ell^T w) \prod_{i=1}^{k} \frac{\ell_i}{\exp(\ell_i r_i)} \mathbb{1}_{\mathscr{W}_{r_i}}(w_i), \quad (14)$$

and

$$\mathscr{W}_{r_i} = \begin{cases} (-\infty, r_i] & \text{if } \ell_i > 0 \\ [r_i, \infty) & \text{if } \ell_i < 0 \end{cases} \tag{15}$$

where $\ell, r \in \mathbb{R}^k$ k-real value. Otherwise, for all $\ell_i = 0$ the density $p(w)$ becomes a multivariate uniform distribution with $\mathscr{W}_{r_i} = [-r_i, r_i]$. Given an $\ell \in \mathbb{L}$, it can be shown that the following set of prior distributions (c.f. [3, th. 4.6])

$$\mathscr{M}^w = \left\{ w \in \mathscr{W} \mid p(w) \propto \exp(\ell^T w), \ell = [\ell_1, \ldots, \ell_k]^T \in \mathbb{L} \right\}, \tag{16}$$

satisfies (P1)-(P4) properties as well as conjugacy between the likelihood and the set of posterior distributions.

Since our g.p.d. $\mathbb{P}_{X|y=m_k}$ given by Equation (10) belongs to $\mathscr{FE}xp$, we can use the set of prior distributions $\mathscr{M}^\mu$ of Equation (16) in order to get a set of posterior distributions $\mathscr{M}^\mu_n$ having the same functional form ($\mathscr{FE}xp$) [5, §5.2]:

$$\mathscr{M}^\mu_n = \left\{ \mu \,|\, \bar{\boldsymbol{x}}_n, \ell \propto \mathscr{N}\left(\frac{\ell + n\bar{\boldsymbol{x}}_n}{n}, \frac{1}{n}\widehat{\Sigma} \right) \right\} \tag{17}$$

where $\bar{\boldsymbol{x}}_n = \frac{1}{n}\sum_{i=1}^n x_i$ and $\ell \in \mathbb{L}$. Using posterior expectations, we can estimate the lower and upper values of the unknown μ parameters, so for every dimension $i \in \{1, \ldots, d\}$:

$$\inf_{\mathscr{M}^\mu_n} \mathbb{E}[\mu_i \,|\, \bar{\boldsymbol{x}}_n, \ell] = \underline{\mathbb{E}}[\mu_i \,|\, \bar{\boldsymbol{x}}_n, \ell] = \frac{-c_i + n\bar{\boldsymbol{x}}_n}{n} \tag{18}$$

$$\sup_{\mathscr{M}^\mu_n} \mathbb{E}[\mu_i \,|\, \bar{\boldsymbol{x}}_n, \ell] = \overline{\mathbb{E}}[\mu_i \,|\, \bar{\boldsymbol{x}}_n, \ell] = \frac{c_i + n\bar{\boldsymbol{x}}_n}{n} \tag{19}$$

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

$$\mathbb{G}_{m_k} = \left\{ \widehat{\mu}_{m_k} \in \mathbb{R}^d \,\middle|\, \begin{array}{l} \widehat{\mu}_{i,m_k} \in \left[\dfrac{-c_i + n_{m_k}\bar{\boldsymbol{x}}_{i,n_{m_k}}}{n_{m_k}}, \dfrac{c_i + n_{m_k}\bar{\boldsymbol{x}}_{i,n_{m_k}}}{n_{m_k}} \right] \\ \quad \forall i = \{1, \ldots, d\},\ c_i > 0 \end{array} \right\}. \tag{20}$$

Remark 4 *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., $\hat{\mu} = \frac{\ell + n\bar{\boldsymbol{x}}_n}{n} \xrightarrow[n\to\infty]{} \bar{\boldsymbol{x}}_n$, and will become the asymptotic estimator of the precise Gaussian distribution.*

On the basis of posterior estimator $\mathbb{G}_{m_k}$ previously calculated, we can write the set of conditional probability distributions $\mathscr{P}_{X|y=m_k}$ for every label $m_k \in \mathscr{K}$ as follows:

$$\mathscr{P}_{X|y=m_k} = \left\{ \mathbb{P}_{X|y=m_k} \,\middle|\, \mathbb{P}_{X|y=m_k} \sim \mathscr{N}(\mu_{m_k}, \widehat{\Sigma}_{m_k}), \mu_{m_k} \in \mathbb{G}_{m_k} \right\} \tag{21}$$

In what follows, we study how we can incorporate the set of distributions $\mathscr{P}_{X|y=m_k}$ in Gaussian discriminant analysis, using maximality (Definition 2) to get our (possibly) *imprecise* classification.

4. Imprecise Classification with $\mathscr{L}_{0/1}$ Loss Function

We first present our approach to make cautious classification by using sets of conditional distribution given by Equation (21) and obtained from a near-ignorance model. Using the maximality criterion, to know whether $m_a \succ_M m_b$, we need to solve

$$\inf_{\mathbb{P}_Y \in \mathscr{P}_Y} \inf_{\substack{\mathbb{P}_{X|m_a} \in \mathscr{P}_{X|m_a} \\ \mathbb{P}_{X|m_b} \in \mathscr{P}_{X|m_b}}} \big[P(X = \boldsymbol{x}|y = m_a)P(y = m_a) - P(X = \boldsymbol{x}|y = m_b)P(y = m_b) \big] > 0 \tag{22}$$

where 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 (22) as follows (cf. [25, eq. 4.3]):

$$\inf_{\mathbb{P}_Y \in \mathscr{P}_Y} \big[\underline{P}(X = \boldsymbol{x}|y = m_a)P(y = m_a) - \overline{P}(X = \boldsymbol{x}|y = m_b)P(y = m_b) \big] > 0 \tag{23}$$

where $\underline{P}$ ($\overline{P}$) is the infimum (supremum) conditional probability. Also, applying Assumption 2, where every $\widehat{\pi}_y > 0$, Equation (23) is reduced to finding the two values

$$\underline{P}(\boldsymbol{x}|y = m_a) = \inf_{\mathbb{P}_{X|m_a} \in \mathscr{P}_{X|m_a}} P(\boldsymbol{x}|y = m_a), \tag{24}$$

$$\overline{P}(\boldsymbol{x}|y = m_b) = \sup_{\mathbb{P}_{X|m_b} \in \mathscr{P}_{X|m_b}} P(\boldsymbol{x}|y = m_b). \tag{25}$$

As $\mathscr{P}_{X|y=m_k}$ is a set of Gaussian distributions, the solutions of Equations (24) and (25) are respectively obtained for the following values of the means

$$\underline{\mu}_{m_a} = \underset{\mu_{m_a} \in \mathbb{G}_{m_a}}{\arg\inf} -\frac{1}{2}(\boldsymbol{x} - \mu_{m_a})^T \widehat{\Sigma}^{-1}_{m_b}(\boldsymbol{x} - \mu_{m_a}), \tag{26}$$

$$\overline{\mu}_{m_b} = \underset{\mu_{m_b} \in \mathbb{G}_{m_b}}{\arg\sup} -\frac{1}{2}(\boldsymbol{x} - \mu_{m_b})^T \widehat{\Sigma}^{-1}_{m_b}(\boldsymbol{x} - \mu_{m_b}), \tag{27}$$

where $\widehat{\Sigma}^{-1}_{m_b}$ is the inverse of the covariance matrix (Assumption 3). Depending on the internal structure of the *precise* covariance matrix $\hat{\Sigma}_{cat_k}$, solving (26) and (27) may be more or less computationally challenging. Here we explore the most general cases, and omit the cases where $\hat{\Sigma}_{m_k}$ is a diagonal matrix due to space limitations.

In particular, we will explore the two following cases:

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 cases where there exists *collinearity* or *multicollinearity* across features of covariance matrices, $\hat{\Sigma}_{m_k}$ will not be invertible, in which case we use the *singular value decomposition* (SVD) method for computing the *pseudo-inverse* of the covariance matrix. Before studying the computational issues of IQDA and ILDA, 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 4, 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 4(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 of training data.

We also drew (precise) mean of each group, i.e. μ_{m_a} and μ_{m_b}, as solid points in the centre of each square, and a black dot ($\bullet$) representing a new unlabelled instance $\boldsymbol{x}$ as well as positions of solutions of Equations (26) and (27). In the Figure 4(b), we observe an area of uncertainty in which we would predict both classes (in purple on the figure), that is generated by the imprecise mean and the maximality criterion.

Let us now discuss the problem of solving Equations (27) and (26). Expressing $\mathbb{G}_{m_b}$ as constraints, the solution $\overline{\mu}_{m_b}$ of (27) can be written as

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

This optimisation problem is well-known as a box-constraint quadratic program (BQP) [9], 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 [14]. Computing an optimal global solution of this convex optimisation problem in polynomial time is easy using modern libraries (e.g. CvxOpt python library).

Finding $\underline{\hat{\mu}}_{m_a}$ in Equation (26) is much more difficult, as one seeks the optimal value

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

That comes down to maximizing a convex function over box-constraints $\mathbb{G}_{m_a}$, which is known to be NP-Hard [17]. To solve it, we use a brand-and-bound (B&B) algorithm [7, 23], that employs a finite branching based on the first-order Karush-Kuhn-Tucker[2] conditions and polyhedral semidefinite relaxation in each node of the B&B tree (more details in [7]).

5. Experiments Setting

In this section, we provide first experimental results to evaluate the performance of our two 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:

$$\mathbb{L}' = \left\{\ell \in \mathbb{R}^k : \ell_i \in [-c, c], c > 0, i = \{1, \ldots, d\}\right\}. \quad (28)$$

In order to fix a value of c, there exists different approaches already mentioned in Section 4.3. of [3]. One can for example rely on the rate of convergence of the lower and upper posterior expectations [22]:

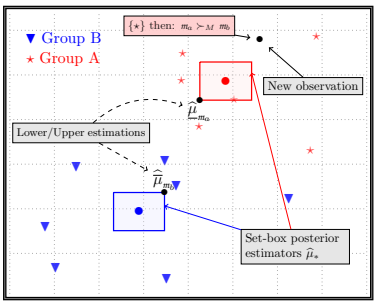
$$\forall i \quad \left(\overline{E}[\mu_i | \overline{\boldsymbol{x}}_n, \ell] - \underline{E}[\mu_i | \overline{\boldsymbol{x}}_n, \ell]\right) = \frac{2c}{n} \xrightarrow[n \to \infty]{} 0 \qquad (29)$$

meaning that for small values of c, we would reach a faster convergence of Equation (29) to a *precise* posterior inference (as precise models). A value of $c \leq 0.75$ is recommended by [3, §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 10-folds on the training samples. A typical empirical evolution of the accuracy measures used in the next sections with the value of c is shown in Figure 5(c).

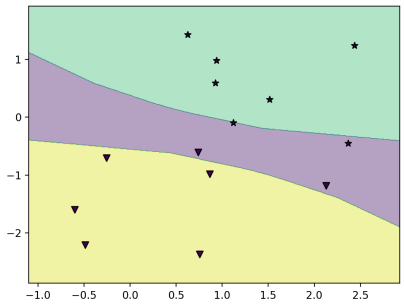
5.2. Data Sets and Experimental Setting

We perform experiments on 9 data sets issued from UCI machine repository [11](cf. Table 1), following 10×10-fold cross-validation procedure. We aim to compare the performance of our imprecise Gaussian classifier model approach with the existing precise models, i.e. Linear and Quadratic Discriminant Analysis (resp. LDA and QDA).

2. Also known as KKT, this one allows to solve problems of optimisation subject to non-linear constraints on the form of inequalities.



(*a*) Upper/lower estimation of μ_{m_k}



(*b*) Imprecise decision area

Figure 4: *Imprecise boundary area and estimation*. Figure 4(*a*) shows an example of imprecise estimation mean μ_*, and Figure 4(*b*) shows an imprecise decision area of color purple where the subset $\widehat{Y}=\{m_a, m_b\}$ of labels is the imprecise decision dominating all the others (i.e. $\{a\}$ or $\{b\}$).

#	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	dermatology	385	34	6
f	vehicle	846	18	4
g	vowel	990	10	11
h	wine-quality	1599	11	6
i	wall-following	5456	24	4

Table 1: Data sets used in the experiments

Comparing indeterminate predictions given in the form of a subset of plausible labels $\widehat{Y}$ against just one plausible label $\widehat{y}$ is a difficult issue 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 unique, possibly risky ones. 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 [26], called *utility-discounted accuracy*, which makes it possible to reward the imprecision in a more or less strong way. It is written as follows:

$$u(y,Y) = \begin{cases} 0 & \text{if } y \notin Y, \\ \frac{\alpha}{|Y|} - \frac{\alpha-1}{|Y|^2} & \text{else.} \end{cases} \tag{30}$$

[26] 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 [24]). 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 for rightful cautiousness.

5.3. Experimental Results

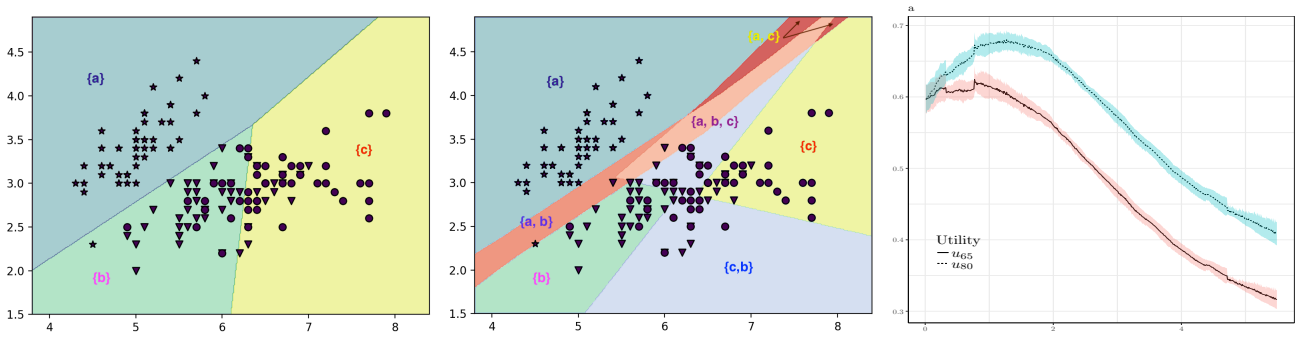
The average results obtained according to u_{65} and u_{80} utilities, and the average execution time to predict the label of a new unlabeled instance are shown in Table 2.

	LDA	**ILDA**		**QDA**	**IQDA**		**Avg.**
#	acc.	u_{80}	u_{65}	acc	u_{80}	u_{65}	**time (sec.)**
a	97.96	**98.38**	97.16	97.29	**98.08**	97.13	0.56
b	98.85	**98.99**	98.95	99.03	**99.39**	99.09	1.49
c	94.61	**94.56**	94.05	89.43	**91.77**	88.90	12.14
d	96.35	**96.59**	96.51	94.64	**95.20**	94.72	1.50
e	96.58	**97.06**	96.94	82.47	**84.24**	84.05	19.24
f	77.96	**81.98**	79.59	85.07	**87.96**	86.13	3.10
g	60.10	**67.45**	62.41	87.83	**89.96**	88.40	4.95
h	59.25	**65.83**	60.31	55.62	**65.85**	60.36	34.85
i	67.96	**71.34**	66.65	65.87	**71.79**	69.75	10.77
avg.	83.68	**86.05**	84.03	80.34	**87.16**	85.33	10.1

Table 2: Average utility-discounted accuracies (%)

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, wine-quality, wall-following and vowel data sets. 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.

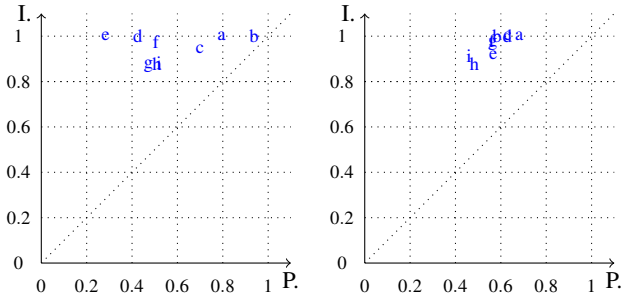
In order to highlight the major role of cautiousness of an imprecise classifier model, we show in Figure 5(*b*) how, in the IRIS data set, our ILDA model creates different areas of decision boundaries (not to be confused with rejection area), where each area represent a different combination of subset of labels $\widehat{Y} \subseteq \mathscr{K}$, in contrast to precise classifier model (LDA), in Figure 5(*a*), where it creates one area for each distinct label. Besides, in Figure 5(*c*), we show the evolution of utility-discounted accuracy (i.e. u_{65} and u_{80}), with a standard deviation calculated by cross-validation 10-



(*a*) Precise boundary decision (*b*) Imprecise decision area (*c*) Evolution c w.r.t. u_{65} and u_{80}

Figure 5: Figure 5(*a*) shows how a precise model divides the instance space in three single different zones by label (i.e $\{a\},\{b\},\{c\}$), the figure 5(*b*) shows how an imprecise 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 5(*c*) shows prediction performance of ILDA model w.r.t. utility-discount accuracy and c tuning parameter on vowel dataset.

fold 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 and non-informative with respect to our attitude towards cautiousness (modelled through utility (30)).



(*a*) ILDA & LDA (*b*) IQDA & QDA

Figure 6: (a) Correctness of the **I**mprecise LDA in the case of abstention versus accuracy of the **P**recise LDA. (b) Correctness of the **I**mpreciseQLDA in the case of abstention versus accuracy of the **P**recise QDA. 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 where the *precise* classifier makes an usual high amount of mistakes. In Figure 6, we verify that our two imprecise classifiers follow this desirable behaviour on most data sets, for the u_{80} measures (conclusions for the u_{65} are similar). Figure 6(*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 method is given by Figure 6(*b*). We notice that on those hard instances where the precise classifiers have a quite lower average accuracy than the global one (i.e., make more mistakes than on the rest of the data), our imprecise classifiers successfully overcome them, getting the ground-truth value into partial predictions (>80%). This is especially true for the linear case, where for the dermatology data set (*e*), the accuracy on the imprecisely classifier instances drop to 30% for the precise classifier (with a global average of over 96%!), while the imprecise classifier always include the true class. Our approach therefore seems to be able to well robustify the very simple, linear decision frontiers of the ILDA models. For instance, Figure 5(*b*) shows that we go from linear to piecewise linear frontiers when going imprecise, showing that the approach does not simply amount to add a rejection threshold when probabilities of classes are too similar, as in this latter case the decision frontiers remain linear.

6. Conclusions

In this paper, we have proposed two new (cautious) imprecise classifiers, which generalize the well-known Linear and Quadratic discriminant classifiers, with the purpose of getting cautious inference in case of insufficient evidence in the available information (i.e. datasets).

In future works, we intend to continue exploring other approaches as well as the complexity of the associated inferences. For instance, (1) we will study the impacts of considering a diagonal structure of the covariance matrix, which would simplify the optimisation step but will reduce the model expressivity, (2) we will release Assumption 2 and consider instead a set of marginals, in order to have an idea of their influence in case of imbalanced datasets and finally, (3) we will consider using a generic loss function in order to generalize *zero-one* loss function (i.e. $\mathscr{L}_{0/1}$).

Another issue that we leave open is the one of estimating an imprecise covariance matrix that makes computations affordable, i.e., whose inverses correspond to an easy-to-estimate and easy-to-deal with convex space of positive definite matrices.

Acknowledgments

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).

References

[1] Thomas Augustin, Frank PA Coolen, Gert de Cooman, and Matthias CM Troffaes. *Introduction to imprecise probabilities*. John Wiley & Sons, 2014.

[2] Alessio Benavoli and Branko 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*, pages 1–8. IEEE, 2011.

[3] Alessio Benavoli and Marco Zaffalon. Prior near ignorance for inferences in the k-parameter exponential family. *Statistics*, 49(5):1104–1140, 2014.

[4] James O Berger. *Statistical decision theory and Bayesian analysis; 2nd ed.* Springer Series in Statistics. Springer, New York, 1985.

[5] José M Bernardo and Adrian FM Smith. *Bayesian Theory*. John Wiley & Sons Ltd., 2000.

[6] Ulisses M. Braga-Neto and Edward R. Dougherty. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3):374–380, 2004.

[7] Samuel Burer and Dieter Vandenbussche. Globally solving box-constrained nonconvex quadratic programs with semidefinite-based finite branch-and-bound. *Computational Optimization and Applications*, 43(2):181–195, 2009.

[8] Lori A. Dalton and Mohammadmahdi R. Yousefi. On optimal bayesian classification and risk estimation under multiple classes. *EURASIP Journal on Bioinformatics and Systems Biology*, 2015(1):8, Oct 2015. ISSN 1687-4153.

[9] Pasquale L De Angelis, Panos M Pardalos, and Gerardo Toraldo. Quadratic programming with box constraints. In *Developments in global optimization*, pages 73–93. Springer US, 1997.

[10] Arthur P Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232, 1968.

[11] A. Frank and A. Asuncion. UCI machine learning repository, 2010. URL http://archive.ics.uci.edu/ml.

[12] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Springer New York Inc., 2001.

[13] Radu Herbei and Marten H Wegkamp. Classification with reject option. *Canadian Journal of Statistics*, 34 (4):709–721, 2006.

[14] CR Johnson. Positive definite matrices. *The American Mathematical Monthly*, 77(3):259–264, 1970.

[15] Rob Kitchin and Tracey P Lauriault. Small data in the era of big data. *GeoJournal*, 80(4):463–475, 2015.

[16] Isaac Levi. *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. MIT press, 1983.

[17] Panos M Pardalos and Stephen A Vavasis. Quadratic programming with one negative eigenvalue is np-hard. *Journal of Global Optimization*, 1(1):15–22, 1991.

[18] Christian Robert. *Le choix bayésien: Principes et pratique*. Springer Paris, 2005.

[19] Robin Senge, Stefan Bösner, Krzysztof Dembczyński, Jörg Haasenritter, Oliver Hirsch, Norbert Donner-Banzhoff, and Eyke Hüllermeier. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255: 16–29, 2014.

[20] Samuel James Taylor. *Introduction to measure and integration*. CUP Archive, 1973.

[21] Matthias CM Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17–29, 2007.

[22] P. Walley. *Statistical reasoning with imprecise Probabilities*. Chapman and Hall, 1991.

[23] Wei Xia, Juan Vera, and Luis F Zuluaga. Globally solving non-convex quadratic programs via linear integer programming techniques. *arXiv preprint arXiv:1511.02423*, 2015.

[24] Gen Yang, Sébastien Destercke, and Marie-Hélène Masson. Cautious classification with nested dichotomies and imprecise probabilities. *Soft Computing*, 21(24):7447–7462, 2017.

[25] Marco Zaffalon. The naive credal classifier. *Journal of statistical planning and inference*, 105(1):5–21, 2002.

[26] Marco Zaffalon, Giorgio Corani, and Denis Mauá. Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning*, 53(8):1282–1301, 2012.