Uncertainty Quantification in Logistic Regression using Random Fuzzy Sets and Belief Functions

Thierry Denœux

Uncertainty Quantification in Logistic Regression using Random Fuzzy Sets and Belief Functions

Thierry Denœux[a,b]

[a] *Université de technologie de Compiègne, CNRS UMR 7253 Heudiasyc, Compiègne, France*
[b] *Institut universitaire de France, Paris, France*

Abstract

Evidential likelihood-based inference is a new approach to statistical inference in which the relative likelihood function is interpreted as a possibility distribution. By expressing new data as a function of the parameter and a random variable with known probability distribution, one then defines a random fuzzy set and an associated predictive belief function representing uncertain knowledge about future observations. In this paper, this approach is applied to binomial and multinomial regression. In the binomial case, the predictive belief function can be computed by numerically integrating the possibility distribution of the posterior probability. In the multinomial case, the solution is obtained by a combination of constrained nonlinear optimization and Monte Carlo simulation. In both cases, computations can be considerably simplified using a normal approximation to the relative likelihood. Numerical experiments show that decision rules based on predictive belief functions make it possible to reach lower error rates for different rejection rates, as compared to decisions based on posterior probabilities.

Keywords: Dempster-Shafer theory, evidence theory, possibility distribution, statistical inference, classification, machine learning.

1. Introduction

The Dempster-Shafer (DS) theory of belief functions, introduced by Dempster [7] and Shafer [48], is a generalization of Bayesian reasoning making it possible to represent and reason with weak evidence that could not be adequately represented by probability distributions [14]. The two main features of DS theory are (1) the use of belief functions for representing evidence, and (2) a mechanism for combining independent items of evidence, known as *Dempster's rule of combination* [20].

In machine learning, DS theory has been applied to clustering [1, 18], classification [10, 34] and partially supervised learning [45, 22]. In classification, an important direction of research has been to design *evidential classifiers* quantifying classification uncertainty using

Email address: Thierry.Denoeux@utc.fr (Thierry Denœux)

belief functions [12, 51]. Thanks to the great flexibility of DS theory, evidential classifiers are able to distinguish between *aleatory* uncertainty, arising from the random data generation mechanism on the one hand, and *epistemic* uncertainty due to insufficient data on the other hand.

Logistic regression is one of the most widely used classification techniques [33]. It is based on a discriminative model representing the logarithms of probability ratios for different classes as linear combinations of predictors, resulting in a linear classifier. The coefficients are fitted by likelihood maximization using the Newton-Raphson algorithm. One of the main reasons for the popularity of logistic regression is that it produces interpretable results; in particular, the contribution of each individual predictor can be assessed by testing the significance of the corresponding coefficient. However, a logistic regression classifier computes point estimates of posterior class probabilities without accounting for epistemic uncertainty. Alternatively, Bayesian logistic regression assumes a prior probability distribution on the regression coefficients, and computes the posterior distribution using Markov Chain Monte Carlo techniques [31]. The predictive posterior class probabilities then account for both random and epistemic uncertainty. However, a drawback of Bayesian inference is its reliance on precise prior probabilities, an unreasonable assumption in case of complete ignorance [48, 52].

In this paper, we investigate another approach to logistic regression based on the theory of belief functions, called *evidential logistic regression*. This approach is based on our recent work on statistical inference using epistemic random fuzzy sets [17, 19, 24]. As we will see, our evidential approach boils down to the Bayesian approach when prior probabilities are provided, but it can be used in the absence of prior information, or with weaker forms of prior information. Our study extends previous work by Xu *et al.* [54] and Minary *et al.* [42], which was limited to binary logistic regression. Specifically, the contributions of the present paper are the following:

1. We show that the amount of computation in evidential binomial regression can be reduced using a normal approximation to the relative likelihood function;
2. We extend the method to multinomial classification using both exact and approximate calculations using, again, a normal approximation to the relative likelihood function;
3. Using numerical experiments, we show that the predictive belief functions computed by evidential logistic regression have better predictive performance (as measured by error-reject curves) than the estimated posterior probabilities computed by classical logistic regression.

We can remark that the approach investigated in this paper differs from that presented in [16], in which we showed that the operations performed in logistic regression and in the softmax layer of neural network classifiers can be interpreted as the combination of elementary Dempster-Shafer mass functions resulting in a *latent* belief function. This previous analysis did not consider sampling or epistemic uncertainty, which is our main focus here. A brief comparison between the two approaches will be presented in Section 5.3.

The rest of the paper is organized as follows. The necessary background about possibility theory, random fuzzy sets and evidential likelihood-based inference is first recalled in Section

2. Binomial and multinomial logistic regression are then addressed, respectively, in Sections 3 and 4. Experimental results are reported in Section 5. Finally, Section 6 summarizes the main findings of the paper and opens up some perspectives.

2. Background

In this section, we briefly introduce the theoretical background needed to understand the rest of the paper. Basic notions about possibility theory and epistemic random fuzzy sets (RFSs) are first recalled, respectively, in Sections 2.1 and 2.2. Evidential likelihood-based inference is then summarized in Section 2.3.

2.1. Fuzzy sets and possibility theory

Possibility theory, initiated by Zadeh in [57], is a formalism for uncertain reasoning based on the representation of partial information about variables of interest by flexible constraints (see [21] for a recent account). It is intimately related to the notion of fuzzy set [55]. Formally, a *fuzzy subset* of a set Θ can be identified to a mapping $\widetilde{F} : \Theta \to [0, 1]$. Each number $\widetilde{F}(\theta)$ is interpreted as a "degree of membership" of element θ in $\widetilde{F}$, seen as a set with unsharp boundaries. The *height* of $\widetilde{F}$ is its supremum; it is denoted by $\mathsf{hgt}(\widetilde{F}) = \sup_{\theta\in\Theta} \widetilde{F}(\theta)$. If $\mathsf{hgt}(\widetilde{F}) = 1$, $\widetilde{F}$ is said to be *normal*. For any $\alpha \in [0, 1]$, the α-cut of $\widetilde{F}$ is the set

$$^{\alpha}\widetilde{F} = \{\theta \in \Theta : \widetilde{F}(\theta) \geq \alpha\}.$$

Extension principle. Let f be a mapping from Θ to some set Λ. Zadeh's extension principle [56] makes it possible to extend f to fuzzy subsets of Θ. The image by f of a fuzzy subset $\widetilde{F}$ of Θ is the fuzzy subset $f(\widetilde{F})$ of Λ defined by

$$f(\widetilde{F})(\lambda) = \sup_{\{\theta\in\Theta: f(\theta)=\lambda\}} \widetilde{F}(\theta). \tag{1}$$

If a variable $\boldsymbol{\theta}$ taking values in Θ is constrained by $\widetilde{F}$, $f(\boldsymbol{\theta})$ is, thus, constrained by $f(\widetilde{F})$.

Possibility and necessity measures. Let $\boldsymbol{\theta}$ be a variable taking values in Θ. Assume that we receive a piece of evidence telling us that "$\boldsymbol{\theta}$ is $\widetilde{F}$", where $\widetilde{F}$ is a normal fuzzy subset of Θ. This evidence induces a *possibility measure* $\Pi_{\widetilde{F}}$ from 2^{Θ} to $[0, 1]$ defined by

$$\Pi_{\widetilde{F}}(B) = \sup_{\theta\in B} \widetilde{F}(\theta), \tag{2}$$

for all $B \subseteq \Theta$. The number $\Pi_{\widetilde{F}}(B)$ is interpreted as the degree of possibility that $\boldsymbol{\theta} \in B$, given that $\boldsymbol{\theta}$ is $\widetilde{F}$ [57]. The corresponding *possibility distribution* is the mapping $\pi_{\widetilde{F}} : \Theta \to [0, 1]$ defined by

$$\pi_{\widetilde{F}}(\theta) = \Pi_{\widetilde{F}}(\{\theta\}) = \widetilde{F}(\theta).$$

It is identical to $\widetilde{F}$: the degree of possibility that $\boldsymbol{\theta} = \theta$ given the flexible constraint "$\boldsymbol{\theta}$ is $\widetilde{F}$" is equal to the degree of membership of θ to fuzzy set $\widetilde{F}$. The dual *necessity measure* is defined as

$$N_{\widetilde{F}}(B) = 1 - \Pi_{\widetilde{F}}(B^c) = \inf_{\theta\notin B} \left[1 - \widetilde{F}(\theta)\right], \tag{3}$$

where B^c denotes the complement of B in Θ. It is easy to show that $N_{\widetilde{F}}$ is completely monotone and is, thus a belief function, while $\Pi_{\widetilde{F}}$ is the dual plausibility function [28].

Conjunctive combination of possibility distributions. Assume that we receive two independent pieces of information telling us that "$\boldsymbol{\theta}$ is $\widetilde{F}$" and "$\boldsymbol{\theta}$ is $\widetilde{G}$", where $\widetilde{F}$ and $\widetilde{G}$ are two fuzzy subsets of Θ. The conjunctive combination of these two pieces of evidence requires some notion of intersection between fuzzy sets. As reviewed in [27], the intersection operation can be extended to fuzzy sets using triangular norms (or t-norms for short). The most common choices are the minimum and product t-norms originally proposed by Zadeh [55]. Given a t-norm $\top$, the corresponding normalized intersection of two fuzzy subsets $\widetilde{F}$ and $\widetilde{G}$ of Θ such that $\sup_{\theta'} \widetilde{F}(\theta') \top \widetilde{G}(\theta') > 0$ is the normal fuzzy subset

$$(\widetilde{F} \cap^*_\top \widetilde{G})(\theta) = \frac{\widetilde{F}(\theta) \top \widetilde{G}(\theta)}{\sup_{\theta'} \widetilde{F}(\theta') \top \widetilde{G}(\theta')}. \tag{4}$$

The product is the only t-norm for which this operation is associative [23]. The normalized product-intersection operator will be denoted by $\odot$.

Gaussian fuzzy numbers and vectors. A *Gaussian fuzzy vector (GFV)* is a normal fuzzy subset $\widetilde{F}$ of $\mathbb{R}^p$ (with $p \geq 1$) such that

$$\widetilde{F}(x) = \exp\left(-\frac{1}{2}(x-m)^T \boldsymbol{H}(x-m)\right)$$

for all $x \in \mathbb{R}^p$, where $m \in \mathbb{R}^p$ is the mode of $\widetilde{F}$, and $\boldsymbol{H} \in \mathbb{R}^{p \times p}$ is a symmetric and positive semidefinite precision matrix. We write $\widetilde{F} \sim \mathsf{GFV}(m, \boldsymbol{H})$. It can easily be shown [43] that the family of GFVs is closed under the normalized product intersection. More precisely, assuming $\boldsymbol{H}_1 + \boldsymbol{H}_2$ to be positive definite, we have

$$\mathsf{GFV}(m_1, \boldsymbol{H}_1) \odot \mathsf{GFV}(m_2, \boldsymbol{H}_2) = \mathsf{GFV}(m_{12}, \boldsymbol{H}_{12})$$

with $m_{12} = (\boldsymbol{H}_1 + \boldsymbol{H}_2)^{-1}(\boldsymbol{H}_1 m_1 + \boldsymbol{H}_2 m_2)$ and $\boldsymbol{H}_{12} = \boldsymbol{H}_1 + \boldsymbol{H}_2$. When $p = 1$, a GFV boils down to a *Gaussian fuzzy number* (GFN) and we write $\mathsf{GFN}(m, h)$ with $m \in \mathbb{R}$ and $h \geq 0$.

The following proposition states that the image of a GFV by a linear mapping is still a GFV. This result will be used extensively in Sections 3 and 4.

Proposition 1. *Let $\beta \in \mathbb{R}^p$ be a p-dimensional real vector constrained by a possibility distribution $\widetilde{\beta} \sim \mathsf{GFV}(m, \boldsymbol{H})$ with mode $m \in \mathbb{R}^p$ and positive definite precision matrix $\boldsymbol{H} \in \mathbb{R}^{p \times p}$. Let $\boldsymbol{U} \in \mathbb{R}^{q \times p}$ be a real matrix of rank $q \leq p$, and $Z = \boldsymbol{U}\beta \in \mathbb{R}^q$. The possibility distribution $\widetilde{Z}$ of Z verifies*

$$\widetilde{Z} \sim \mathsf{GFV}\left(\boldsymbol{U}m, (\boldsymbol{U}\boldsymbol{H}^{-1}\boldsymbol{U}^T)^{-1}\right).$$

Furthermore, the most plausible value of β subject to the constraint $\boldsymbol{U}\beta = z$ is

$$\beta^* = m + \boldsymbol{H}^{-1}\boldsymbol{U}^T(\boldsymbol{U}\boldsymbol{H}^{-1}\boldsymbol{U}^T)^{-1}(z - \boldsymbol{U}m).$$

Proof. See Appendix A. □

Proposition 1 makes it possible, in particular, to compute marginals of GFVs. Let $m = (m_1, m_2)$, where m_1 and m_2 are two subvectors of m of respective lengths r and s, with $p = r + s$, and consider the corresponding block decomposition of $\boldsymbol{H}$:

$$\boldsymbol{H} = \begin{pmatrix} \boldsymbol{H}_{11} & \boldsymbol{H}_{12} \\ \boldsymbol{H}_{21} & \boldsymbol{H}_{22} \end{pmatrix}.$$

We can write $m_1 = \boldsymbol{U}m$ and $\boldsymbol{H}_{11} = \boldsymbol{U}\boldsymbol{H}\boldsymbol{U}^T$, where $\boldsymbol{U}$ is the matrix of size $r \times p$

$$\boldsymbol{U} = \begin{pmatrix} \mathbf{Id}_r & \mathbf{0}_{r\times s} \end{pmatrix},$$

where $\mathbf{Id}_r$ is the identity matrix of size $r \times r$, and $\mathbf{0}_{r\times s}$ is the null matrix of size $r \times s$. Consider the block decomposition of $\boldsymbol{H}^{-1}$:

$$\boldsymbol{H}^{-1} = \begin{pmatrix} [\boldsymbol{H}^{-1}]_{11} & [\boldsymbol{H}^{-1}]_{12} \\ [\boldsymbol{H}^{-1}]_{21} & [\boldsymbol{H}^{-1}]_{22} \end{pmatrix}.$$

From Proposition 1, the marginal of $\mathsf{GFV}(m, \boldsymbol{H})$ with respect to the first r coordinates is the GFV with mode m_1 and precision matrix $([\boldsymbol{H}^{-1}]_{11})^{-1}$, which using formulas for the inverse of a block matrix [43, p. 46] can be written as

$$([\boldsymbol{H}^{-1}]_{11})^{-1} = \boldsymbol{H}_{11} - \boldsymbol{H}_{12}\boldsymbol{H}_{22}^{-1}\boldsymbol{H}_{21}. \tag{5}$$

This result was stated as Lemma 3 in [19].

2.2. Random fuzzy sets

General definitions. The theory of epistemic RFSs [17, 19] is an extension of DS and possibility theories, in which evidence is represented by RFSs. Mathematically, a RFS is defined as follows [5]. Let $(\Omega, \Sigma_\Omega, P)$ be a probability space, (Θ, Σ_Θ) a measurable space, and $\widetilde{X}$ a mapping from Ω to the set $[0,1]^\Theta$ of fuzzy subsets of Θ. For any $\alpha \in [0,1]$, we define the mapping ${}^\alpha\widetilde{X}$ from Ω to 2^Θ as ${}^\alpha\widetilde{X}(\omega) = {}^\alpha[\widetilde{X}(\omega)]$. We say that $\widetilde{X}$ is a RFS if, for any $\alpha \in [0,1]$, ${}^\alpha\widetilde{X}$ is $\Sigma_\Omega - \Sigma_\Theta$ strongly measurable, i.e., for any $B \in \Sigma_\Theta$,

$$\{\omega \in \Omega : {}^\alpha\widetilde{X}(\omega) \cap B \neq \emptyset\} \in \Sigma_\Omega.$$

Interpretation. In epistemic RFS theory, a RFS is a model of uncertain and fuzzy evidence, in which Θ is a domain of an uncertain variable $\boldsymbol{\theta}$, and Ω is a set of possible interpretations of a piece of evidence about $\boldsymbol{\theta}$. If $\omega \in \Omega$ holds, $\boldsymbol{\theta}$ is constrained by the possibility distribution defined by fuzzy set $\widetilde{X}(\omega)$. We do not know for sure which interpretation is the true one, but our beliefs about the true interpretation are represented by probability measure P.

Belief and plausibility functions. Assuming interpretation ω holds, the degree of possibility that $\boldsymbol{\theta}$ belongs to some set $B \in \Sigma_\Theta$ can be calculated from (2) as

$$\Pi_{\widetilde{X}(\omega)}(B) = \sup_{\theta \in B} \widetilde{X}(\omega)(\theta),$$

and the degree of necessity of the same event is $N_{\widetilde{X}(\omega)}(B) = 1 - \Pi_{\widetilde{X}(\omega)}(B^c)$. Let $Bel_{\widetilde{X}}(B)$ and $Pl_{\widetilde{X}}(B)$ denote, respectively, the *expected necessity* and the *expected possibility* of B:

$$Bel_{\widetilde{X}}(B) = \int_\Omega N_{\widetilde{X}(\omega)}(B) dP(\omega), \tag{6a}$$

$$Pl_{\widetilde{X}}(B) = \int_\Omega \Pi_{\widetilde{X}(\omega)}(B) dP(\omega) = 1 - Bel_{\widetilde{X}}(B^c). \tag{6b}$$

The mappings $B \mapsto Bel_{\widetilde{X}}(B)$ and $B \mapsto Pl_{\widetilde{X}}(B)$, are, respectively, belief and plausibility functions [58, 5].

Combination. The product-intersection rule for combining RFSs, introduced in [17] in the discrete case and in [19] in the general case, is defined as follows. Let us consider two probability spaces $(\Omega_i, \Sigma_{\Omega_i}, P_i)$, $i = 1, 2$ and two RFSs $\widetilde{X}_i : \Omega_i \to [0,1]^\Theta$, $i = 1, 2$ representing independent pieces of evidence about variable $\boldsymbol{\theta}$ taking values in Θ. We define a new probability space as $(\Omega_1 \times \Omega_2, \Sigma_{\Omega_1} \otimes \Sigma_{\Omega_2}, P_{12})$, where $\otimes$ denotes the tensor product of sigma-algebras and P_{12} is the probability measure obtained by conditioning the product measure $P_1 \times P_2$ by the fuzzy set of consistent pairs of interpretations with membership function

$$\widetilde{\Theta}_{12}(\omega_1, \omega_2) = \sup_{\theta \in \Theta} \left(\widetilde{X}_1(\omega_1)(\theta) \cdot \widetilde{X}_2(\omega_2)(\theta) \right).$$

The mapping $\widetilde{X}_1 \oplus \widetilde{X}_2 : \Omega_1 \times \Omega_2 \to [0,1]^\Theta$ such that

$$(\widetilde{X}_1 \oplus \widetilde{X}_2)(\omega_1, \omega_2) = \widetilde{X}_1(\omega_1) \odot \widetilde{X}_2(\omega_2),$$

where $\odot$ denotes the fuzzy set normalized product intersection (4) is called the *orthogonal sum* of $\widetilde{X}_1$ and $\widetilde{X}_2$. The product-intersection operator $\oplus$ is commutative and associative. It generalizes both Dempster's rule for combining belief functions, and the normalized product intersection of possibility measures.

It must be emphasized here that the product intersection $\oplus$ defines an operation on RFSs, and not on belief functions [17, 19]. In particular, any fuzzy subset $\widetilde{F}$ of Θ can be associated with a constant RFS $\widetilde{X}_{\widetilde{F}}$ such that $\widetilde{X}_{\widetilde{F}}(\omega) = \widetilde{F}$ for all $\omega \in \Omega$, or with a random crisp (i.e., nonfuzzy) set $\overline{X}_{\widetilde{F}} : \Omega \to 2^\Theta$, where Ω is the interval $[0,1]$ equipped with the uniform probability measure, such that $\overline{X}_{\widetilde{F}}(\omega) = {}^\omega\widetilde{F}$ for all $\omega \in \Omega$. Random set $\overline{X}_{\widetilde{F}}$ is said to be *consonant* as, for any $(\omega, \omega') \in \Omega^2$, we have $\overline{X}_{\widetilde{F}}(\omega) \subseteq \overline{X}_{\widetilde{F}}(\omega')$ or $\overline{X}_{\widetilde{F}}(\omega') \subseteq \overline{X}_{\widetilde{F}}(\omega)$. It is easy to see that $\widetilde{X}_{\widetilde{F}}$ and $\overline{X}_{\widetilde{F}}$ correspond to the same belief function, i.e., $Bel_{\widetilde{X}_{\widetilde{F}}} = Bel_{\overline{X}_{\widetilde{F}}}$. Yet, given two fuzzy subsets $\widetilde{F}$ and $\widetilde{G}$ of Θ, $\widetilde{X}_{\widetilde{F}} \oplus \widetilde{X}_{\widetilde{G}} = \widetilde{X}_{\widetilde{F} \odot \widetilde{G}}$ is different from $\overline{X}_{\widetilde{F}} \oplus \overline{X}_{\widetilde{G}}$ (the former is still a consonant RFS, while the latter is a RS that is no longer consonant).

2.3. Evidential likelihood-based inference

A theory of statistical inference based on epistemic RFSs was proposed in [17, 19], as an improvement of a previous approach introduced in [13, 36, 37]. It is briefly summarized here. We consider an observed random vector $\boldsymbol{Y}$ with probability density function (pdf) $f_{\boldsymbol{Y}|\boldsymbol{\theta}}$, where $\boldsymbol{\theta} \in \Theta$ is the unknown parameter[1]. The likelihood of any value θ of the parameter after observing $\boldsymbol{Y} = \boldsymbol{y}$ is

$$L(\theta; \boldsymbol{y}) = c f_{\boldsymbol{Y}|\theta}(\boldsymbol{y}),$$

where c is an arbitrary positive constant. Assuming that $\sup_\theta L(\theta; \boldsymbol{y}) < +\infty$, we can define the relative likelihood of θ as

$$\pi_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta) = \frac{L(\theta; \boldsymbol{y})}{\sup_{\theta' \in \Theta} L(\theta'; \boldsymbol{y})} = \frac{L(\theta; \boldsymbol{y})}{L(\widehat{\theta}; \boldsymbol{y})}, \tag{7}$$

where $\widehat{\theta}$ is the maximum likelihood estimate (MLE) of θ. We interpret mapping $\pi_{\theta|\boldsymbol{y}} : \Theta \to [0, 1]$ as a *possibility distribution* over Θ or, equivalently, as the *fuzzy set* of likely values of $\boldsymbol{\theta}$ after observing $\boldsymbol{Y} = \boldsymbol{y}$. It is, thus, a representation of the information about $\boldsymbol{\theta}$ provided by observation $\boldsymbol{y}$. For any $A \subseteq \Theta$, the degrees of plausibility and belief and that $\boldsymbol{\theta} \in A$ after observing $\boldsymbol{y}$ can be computed, respectively, as

$$Pl_{\boldsymbol{\theta}|\boldsymbol{y}}(A) = \sup_{\theta \in A} \pi_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta) \quad \text{and} \quad Bel_{\boldsymbol{\theta}|\boldsymbol{y}}(A) = 1 - Pl_{\pi_{\boldsymbol{\theta}|\boldsymbol{y}}}(A^c). \tag{8}$$

This representation was justified in [13] and [17] as the least committed solution verifying the following two requirements:

R_1**:** Compatibility with Bayesian inference: let P_0 be a prior probability measure on Θ; then, $P_0 \oplus \pi_{\boldsymbol{\theta}|\boldsymbol{y}} = P_{\boldsymbol{\theta}|\boldsymbol{y}}$, where $\oplus$ is the product-intersection operator recalled in Section 2.2, and $P_{\boldsymbol{\theta}|\boldsymbol{y}}$ the Bayesian posterior probability measure on Θ;

R_2**:** Combination of independent observations: let $\boldsymbol{y}$ and $\boldsymbol{y}'$ be independent observations; then, $\pi_{\boldsymbol{\theta}|\boldsymbol{y}} \oplus \pi_{\boldsymbol{\theta}|\boldsymbol{y}'} = \pi_{\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{y}'}$.

Requirement R_1 implies that our approach is an extension of Bayesian inference, in which prior information no longer needs to be assumed. We can also remark that weaker forms of prior information than considered in Bayesian inference such as, for instance, a priori possibility distributions can easily be accommodated in our approach, as will be shown below. Requirement R_2 ensures that $\pi_{\boldsymbol{\theta}|\boldsymbol{y}}$ captures all the information about the parameter provided by $\boldsymbol{y}$: after computing function $\pi_{\boldsymbol{\theta}|\boldsymbol{y}}$, we do not need to store the original data $\boldsymbol{y}$; if a new independent observation $\boldsymbol{y}'$ is made, combining $\pi_{\boldsymbol{\theta}|\boldsymbol{y}}$ with $\pi_{\boldsymbol{\theta}|\boldsymbol{y}'}$ using the product-intersection operator gives us the same result as the one obtained by concatenating the two

[1]We use boldface character $\boldsymbol{\theta}$ for the unknown parameter, and θ for an arbitrary value of $\boldsymbol{\theta}$.

observations[2]. A brief comparison with other evidential approaches to statistical inference, in particular Inferential Models [3][39][40][41] is presented in Appendix B

Normal approximation. Assuming $\ln \pi_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta)$ to be twice differentiable, a tractable approximation of function $\pi_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta)$ can often be obtained by computing a Taylor expansion of its logarithm about the MLE $\widehat{\theta}$ up to the second order [50]:

$$\ln \pi_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta) = \ln \pi_{\boldsymbol{\theta}|\boldsymbol{y}}(\widehat{\theta}) + (\theta - \widehat{\theta})^T \left.\frac{\partial \ln \pi_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta)}{\partial \theta}\right|_{\theta=\widehat{\theta}} + \frac{1}{2}(\theta - \widehat{\theta})^T \left.\frac{\partial^2 \ln \pi_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta)}{\partial \theta \partial \theta^T}\right|_{\theta=\widehat{\theta}} (\theta - \widehat{\theta}) + \cdots$$

The first term on the right-hand side of the above equation is zero by definition, and the second term is zero in the usual case where $\widehat{\theta}$ is a stationary point of $\pi_{\boldsymbol{\theta}|\boldsymbol{y}}$. Neglecting the remaining terms of the Taylor expansion, we get the following approximation

$$\pi_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta) \approx \exp\left[-\frac{1}{2}(\theta - \widehat{\theta})^T \boldsymbol{\mathcal{I}}(\widehat{\theta})(\theta - \widehat{\theta})\right], \tag{9}$$

where $\boldsymbol{\mathcal{I}}(\widehat{\theta})$ is the *observed information matrix* defined as

$$\boldsymbol{\mathcal{I}}(\widehat{\theta}) = -\left.\frac{\partial^2 \ln \pi_{\boldsymbol{\theta}|\boldsymbol{y}}}{\partial \theta \partial \theta^T}\right|_{\theta=\widehat{\theta}} = -\left.\frac{\partial^2 \ln L(\theta; \boldsymbol{y})}{\partial \theta \partial \theta^T}\right|_{\theta=\widehat{\theta}}.$$

As noted in [50], this approximation is usually well verified when $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is an independent sample and n is large.

Prior information. As mentioned above, combination of the likelihood-based possibility distribution (7) with a Bayesian prior using the product-intersection rule yields the Bayesian posterior. However, prior knowledge, when available, is usually vague and, in most cases, there does not seem to be any compelling reason to represent it by a probability distribution. In the robust Bayes approach, prior information is represented by a set of probability distributions, an approach also advocated by Martin [39] in the context of Inferential Models. In [24], I proposed to encode prior knowledge as a RFS, a very general model encompassing Bayesian, possibilistic and vacuous priors as special cases. In particular, a constant RFS, i.e., a possibility distribution is often a simple and convenient model of weak prior information [26]. Let $\pi_{\boldsymbol{\theta}|\text{prior}}$ be a prior possibility distribution on $\boldsymbol{\theta}$. Combining it with the likelihood-based possibility distribution (7) yields the posterior possibility distribution

$$\pi_{\boldsymbol{\theta}|\boldsymbol{y},\text{prior}} = \pi_{\boldsymbol{\theta}|\boldsymbol{y}} \odot \pi_{\boldsymbol{\theta}|\text{prior}},$$

[2]In earlier work [13, 36, 37], the relative likelihood was interpreted as the contour function of a consonant belief function. However, the combination by Dempster's rule of the consonant belief functions induced by two independent samples is not equal to the consonant belief function induced by the union of the two samples, a contradiction that was remarked in [13]. The possibilistic interpretation of the relative likelihood resolves this contradiction.

which can be computed as

$$\pi_{\boldsymbol{\theta}|\boldsymbol{y},\text{prior}} = \frac{L(\theta;\boldsymbol{y})\pi_{\boldsymbol{\theta}|\text{prior}}(\theta)}{\sup_{\theta'\in\Theta} L(\theta';\boldsymbol{y})\pi_{\boldsymbol{\theta}|\text{prior}}(\theta')} = \frac{L(\theta;\boldsymbol{y})\pi_{\boldsymbol{\theta}|\text{prior}}(\theta)}{L(\widehat{\theta}';\boldsymbol{y})\pi_{\boldsymbol{\theta}|\text{prior}}(\widehat{\theta}')},$$

where $\widehat{\theta}'$ is a maximizer of $L(\cdot;\boldsymbol{y})\pi_{\boldsymbol{\theta}|\text{prior}}$. As before, a normal approximation of $\pi_{\boldsymbol{\theta}|\boldsymbol{y},\text{prior}}$ can be obtained by computing a second-order Taylor series expansion of $\ln \pi_{\boldsymbol{\theta}|\boldsymbol{y},\text{prior}}$ about $\widehat{\theta}'$. Alternatively, if the prior possibility distribution is itself Gaussian, i.e., if it is of the form

$$\pi_{\boldsymbol{\theta}|\text{prior}}(\theta) = \exp\left[-\frac{1}{2}(\theta-\theta_0)^T\boldsymbol{H}_0(\theta-\theta_0)\right],$$

it can be combined with the normal approximation (9), resulting in the following approximation:

$$\pi_{\boldsymbol{\theta}|\boldsymbol{y},\text{prior}}(\theta) \approx \exp\left[-\frac{1}{2}(\theta-\widehat{\theta}_1)^T[\boldsymbol{H}_0+\boldsymbol{\mathcal{I}}(\widehat{\theta})](\theta-\widehat{\theta}_1)\right],$$

with

$$\widehat{\theta}_1 = [\boldsymbol{H}_0+\boldsymbol{\mathcal{I}}(\widehat{\theta})]^{-1}\left(\boldsymbol{H}_0\theta_0+\boldsymbol{\mathcal{I}}(\widehat{\theta})\widehat{\theta}\right).$$

In practice, situations in which the data analyst has true and reliable prior knowledge are rather rare. For this reason, no prior knowledge will be assumed in the developments below. The special case of regularization will be briefly addressed in Section 3.1.

Prediction. Let us now consider a prediction problem, where we want to predict the value of a new Y_{new} with sample space $\mathcal{Y}$, whose distribution also depends on $\boldsymbol{\theta}$. We can always define a random variable Y^* with the same distribution as that of Y_{new}, such that

$$Y^* = \varphi(\boldsymbol{\theta}, U), \tag{10}$$

where U is a pivotal random variable with known distribution and sample space $\mathcal{U}$, and φ is a mapping from $\Theta\times\mathcal{U}$ to $\mathcal{Y}$ [37]. We call (10) a *φ-equation*. After observing the data $\boldsymbol{y}$, our knowledge about $\boldsymbol{\theta}$ is represented by the possibility distribution $\pi_{\boldsymbol{\theta}|\boldsymbol{y}}$. By Zadeh's extension principle (1), our knowledge of Y_{new} conditionally on $U=u$ is, thus, represented by the possibility distribution $\pi_{Y_{\text{new}}|\boldsymbol{y},u} = \varphi(\pi_{\boldsymbol{\theta}|\boldsymbol{y}}, u)$ defined as

$$\pi_{Y_{\text{new}}|\boldsymbol{y},u}(y) = \sup_{\{\theta\in\Theta:\varphi(\theta,u)=y\}} \pi_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta) \tag{11}$$

for all $y\in\mathcal{Y}$. The mapping

$$\begin{aligned}\widetilde{Y}: \quad [0,1] &\to [0,1]^{\mathcal{Y}}\\ u &\mapsto \pi_{Y_{\text{new}}|\boldsymbol{y},u}\end{aligned}$$

is, then, a RFS representing statistical evidence about Y_{new}. The corresponding *predictive belief function* $Bel_{\widetilde{Y}}$ and the dual plausibility function $Pl_{\widetilde{Y}}$ can be computed by (6).

We can remark that, if we draw $\boldsymbol{\theta}$ from its posterior distribution $f(\theta|\boldsymbol{y})$ and Y^* using the φ-equation (10), the distribution of Y^* given $\boldsymbol{y}$ is identical to the Bayesian predictive distribution of Y_{new} given $\boldsymbol{y}$. Indeed,

$$f(y^*|\boldsymbol{y}) = \int f(y^*|\theta)f(\theta|\boldsymbol{y})d\theta = \int f(y|\theta)f(\theta|\boldsymbol{y})d\theta = f(y|\boldsymbol{y}).$$

Example 1. *Assume that* $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ *is an independent and identically distributed (iid) sample from the Bernoulli distribution* $B(\boldsymbol{\theta})$*. The fuzzy set of likely values of* $\boldsymbol{\theta}$ *after observing* $\boldsymbol{Y} = \boldsymbol{y}$ *is*

$$\pi_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta) = \left(\frac{\theta}{\widehat{\theta}}\right)^{n\widehat{\theta}} \left(\frac{1-\theta}{1-\widehat{\theta}}\right)^{n(1-\widehat{\theta})},$$

where $\widehat{\theta} = n^{-1}\sum_{i=1}^{n} y_i$ *is the MLE of* $\boldsymbol{\theta}$*. Now, let* $Y_{new} \sim B(\boldsymbol{\theta})$*, independent from* $\boldsymbol{Y}$*; it has the same distribution as*

$$Y^* = \varphi(\theta, U) = \begin{cases} 1 & \text{if } U \leq \theta \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

where U *is a random variable with a standard uniform distribution. For* $U = u$*, the possibility distribution* $\pi_{Y_{new}|\boldsymbol{y},u}$ *defined on* $\mathcal{Y} = \{0, 1\}$ *is*

$$\pi_{Y_{new}|\boldsymbol{y},u}(1) = \sup_{\{\theta\in[0,1]:\varphi(\theta,u)=1\}} \pi_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta) = \begin{cases} 1 & \text{if } u \leq \widehat{\theta} \\ \pi_{\boldsymbol{\theta}|\boldsymbol{y}}(u) & \text{otherwise,} \end{cases}$$

and

$$\pi_{Y_{new}|\boldsymbol{y},u}(0) = \sup_{\{\theta\in[0,1]:\varphi(\theta,u)=0\}} \pi_{\boldsymbol{\theta}|\boldsymbol{y}} = \begin{cases} 1 & \text{if } u \geq \widehat{\theta} \\ \pi_{\boldsymbol{\theta}|\boldsymbol{y}}(u) & \text{otherwise.} \end{cases}$$

The plausibility function of the RFS $\widetilde{Y} : u \mapsto \widetilde{Y}(u) = \pi_{Y_{new}|\boldsymbol{y},u}$ *can be computed using* (6b) *as*

$$\begin{aligned}
Pl_{\widetilde{Y}}(\{1\}) &= \mathbb{E}[\widetilde{Y}(U)(1)] && \text{(13a)}\\
&= \mathbb{E}[\widetilde{Y}(U)(1)|U \leq \widehat{\theta}]\ P(U \leq \widehat{\theta}) + \mathbb{E}[\widetilde{Y}(U)(1)|U > \widehat{\theta}]\ P(U > \widehat{\theta}) && \text{(13b)}\\
&= 1 \times \widehat{\theta} + \frac{\int_{\widehat{\theta}}^{1} \pi_{\boldsymbol{\theta}|\boldsymbol{y}}(u)du}{1-\widehat{\theta}} \times (1-\widehat{\theta}) && \text{(13c)}\\
&= \widehat{\theta} + \int_{\widehat{\theta}}^{1} \pi_{\boldsymbol{\theta}|\boldsymbol{y}}(u)du, && \text{(13d)}
\end{aligned}$$

and

$$\begin{aligned}
Pl_{\widetilde{Y}}(\{0\}) &= \mathbb{E}[\widetilde{Y}(U)(0)] && \text{(14a)}\\
&= \mathbb{E}[\widetilde{Y}(U)(0)|U \leq \widehat{\theta}]\ P(U \leq \widehat{\theta}) + \mathbb{E}[\widetilde{Y}(U)(0)|U > \widehat{\theta}]\ P(U > \widehat{\theta}) && \text{(14b)}\\
&= \frac{\int_{0}^{\widehat{\theta}} \pi_{\boldsymbol{\theta}|\boldsymbol{y}}(u)du}{\widehat{\theta}} \times \widehat{\theta} + 1 \times (1-\widehat{\theta}) && \text{(14c)}\\
&= 1 - \widehat{\theta} + \int_{0}^{\widehat{\theta}} \pi_{\boldsymbol{\theta}|\boldsymbol{y}}(u)du. && \text{(14d)}
\end{aligned}$$

The corresponding predictive mass function $m_{\widetilde{Y}}$ is

$$m_{\widetilde{Y}}(\{0\}) = 1 - Pl_{\widetilde{Y}}(\{1\}) = 1 - \widehat{\theta} - \int_{\widehat{\theta}}^{1} \pi_{\boldsymbol{\theta}|\boldsymbol{y}}(u)du \tag{15a}$$

$$m_{\widetilde{Y}}(\{1\}) = 1 - Pl_{\widetilde{Y}}(\{0\}) = \widehat{\theta} - \int_{0}^{\widehat{\theta}} \pi_{\boldsymbol{\theta}|\boldsymbol{y}}(u)du \tag{15b}$$

$$m_{\widetilde{Y}}(\{0,1\}) = 1 - m_{\widetilde{Y}}(\{0\}) - m_{\widetilde{Y}}(\{1\}) = \int_{0}^{1} \pi_{\boldsymbol{\theta}|\boldsymbol{y}}(u)du. \tag{15c}$$

3. Binomial logistic regression

In this section, we apply the general theory recalled in Section 2.3 to binomial logistic regression. The results reported in [54] (using the consonant interpretation of the relative likelihood) are recovered, and some new results are presented. Estimation of coefficients and posterior probabilities are first addressed, respectively, in Sections 3.1 and 3.2. Prediction is then dealt with in Section 3.3.

3.1. Estimation of coefficients

Model. Let us consider a binary classification problem in which the task is to predict a binary response $Y \in \{0, 1\}$ from p features X_j, $j = 1, \ldots p$. Let $X = (1, X_1, \ldots, X_p)$ denote the extended feature vector of dimension $p + 1$. The conditional probability that $Y = 1$ given $X = x$, denoted by $\theta(x; \beta)$, is assumed to be of the following form,

$$\theta(x; \beta) = F_L(\beta^T x), \tag{16}$$

where $F_L(z) = [1 + \exp(-z)]^{-1}$ is the cumulative distribution function (cdf) of the standard logistic distribution function, and $\beta = (\beta_0, \beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ is a vector of unknown coefficients. We note that other cdfs can be used instead of F_L. For instance, choosing the standard normal cdf Φ gives us the probit regression model. The approach studied in this paper for the case of the widely-used binomial logistic model can easily be transferred to the probit and other models.

Estimation of β. Given n independent observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ of Y with corresponding feature vectors $x_i, \ldots, x_n$, the conditional likelihood is

$$L(\beta) = \prod_{i=1}^{n} \theta(x_i; \beta)^{y_i} [1 - \theta(x_i; \beta)]^{1-y_i}. \tag{17}$$

Let $\widehat{\beta}$ be the MLE of β found by maximizing (17) using an iterative nonlinear optimization algorithm. The possibility distribution of β is given by

$$\pi_{\beta|\boldsymbol{y}}(\beta) = \frac{L(\beta; \boldsymbol{y})}{L(\widehat{\beta}; \boldsymbol{y})}. \tag{18}$$

The observed information matrix can be written as

$$\boldsymbol{\mathcal{I}}(\widehat{\beta}) = \boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X},$$

where $\boldsymbol{X}$ is the $n \times (p+1)$ matrix in which each row i contains feature vector x_i, and $\boldsymbol{V}$ is a $n \times n$ diagonal matrix with general element $\widehat{\theta}_i(1-\widehat{\theta}_i)$, where $\widehat{\theta}_i = \theta(x_i; \widehat{\beta})$ (see [33, page 38]). The normal approximation (9) gives us

$$\pi_{\beta|\boldsymbol{y}}(\beta) \approx \exp\left(-\frac{1}{2}(\beta - \widehat{\beta})^T \boldsymbol{\mathcal{I}}(\widehat{\beta})(\beta - \widehat{\beta})\right). \tag{19}$$

The possibility distribution $\widetilde{\beta}$ is, thus approximated by a GFV with mode $\widehat{\beta}$ and precision matrix $\boldsymbol{\mathcal{I}}(\widehat{\beta})$, which we denote by $\pi_{\beta|\boldsymbol{y}} \sim \mathsf{GFV}(\widehat{\beta}, \boldsymbol{\mathcal{I}}(\widehat{\beta}))$.

Marginalization. For interpretation, it is often useful to examine the possibility distribution of individual coefficients, or groups of coefficients $\beta_J = (\beta_j)_{j \in J}$ for $J \subset \{0, \ldots, p\}$. For instance, hypotheses such as $\beta_j = 0$ or $\beta_J = 0$ can be assessed by computing their degree of possibility, an alternative to frequentist tests of significance [37]. To compute the marginal possibility $\pi_{\beta_J|\boldsymbol{y}}(\beta_J)$, we need to solve the following nonlinear optimization problem:

$$\pi_{\beta_J|\boldsymbol{y}}(\beta_J) = \max_{\beta_{\overline{J}}} \pi_{\beta|\boldsymbol{y}}(\beta), \tag{20}$$

where $\beta_{\overline{J}}$ denotes the subvector of β with components in J removed. Using the normal approximation (19) and the expression (5) for the marginal precision matrix of a GRV, we obtain the following normal approximation of $\pi_{\beta_J|\boldsymbol{y}}(\beta_J)$:

$$\pi_{\beta_J|\boldsymbol{y}}(\beta_J) \approx \exp\left(-\frac{1}{2}(\beta_J - \widehat{\beta}_J)^T (\boldsymbol{\mathcal{I}}_{J,J} - \boldsymbol{\mathcal{I}}_{J,\overline{J}} \boldsymbol{\mathcal{I}}_{\overline{J},\overline{J}}^{-1} \boldsymbol{\mathcal{I}}_{\overline{J},J})(\beta_J - \widehat{\beta}_J)\right), \tag{21}$$

where $\boldsymbol{\mathcal{I}}_{J,\overline{J}}$ is the submatrix of $\boldsymbol{\mathcal{I}}(\widehat{\beta})$ obtained by selecting the rows $i \in J$ and leaving out the columns $j \notin J$, and the other notations $\boldsymbol{\mathcal{I}}_{J,J}$, $\boldsymbol{\mathcal{I}}_{\overline{J},\overline{J}}$ and $\boldsymbol{\mathcal{I}}_{\overline{J},J}$ have similar obvious meanings. Furthermore, as shown in [19, Appendix J], the vector $\beta^*_{\overline{J}}$ maximizing (19) with β_J fixed is

$$\beta^*_{\overline{J}} = \widehat{\beta}_{\overline{J}} - \boldsymbol{\mathcal{I}}_{\overline{J},\overline{J}}^{-1} \boldsymbol{\mathcal{I}}_{\overline{J},J}(\beta_J - \widehat{\beta}_J).$$

This value can be used as a starting point when solving optimization problem (20).

Example 2. *As a first example, let us consider the* `CHDAGE` *dataset used in [33, Chapter 1] and available in the R package* `aplore3` *[2]. The dataset contains the age in years (*`age`*), and presence or absence of evidence of significant coronary heart disease (*`chd`*) for 100 subjects in a hypothetical study of risk factors for heart disease (see Figure 1). In this case, the model is* $\texttt{chd} = \beta_0 + \beta_1 \texttt{age}$*. The exact and approximate possibility distributions on* $\beta = (\beta_0, \beta_1)^T$ *are shown in Figure 2. The marginal possibility distributions of* β_0 *and* β_1 *are shown in Figure 3. We can see that the normal approximation is quite accurate around the MLE, with larger errors occurring farther away from the MLE.*

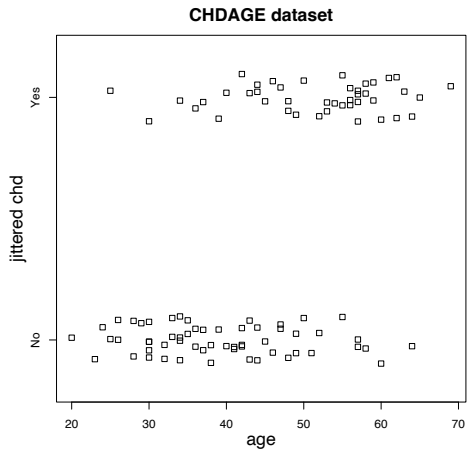


Figure 1: Jitter stripchart of the `CHDAGE` dataset.

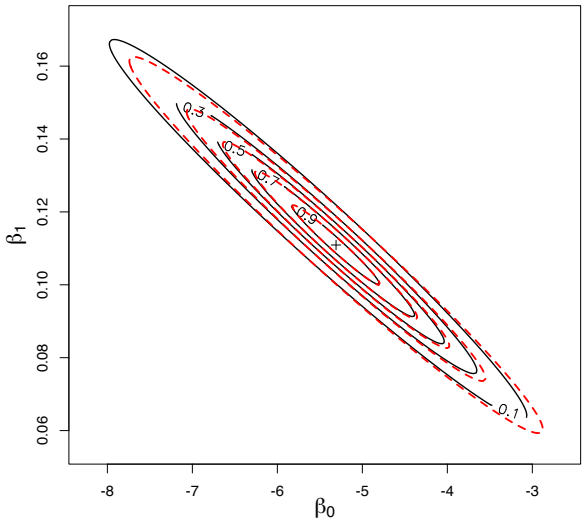
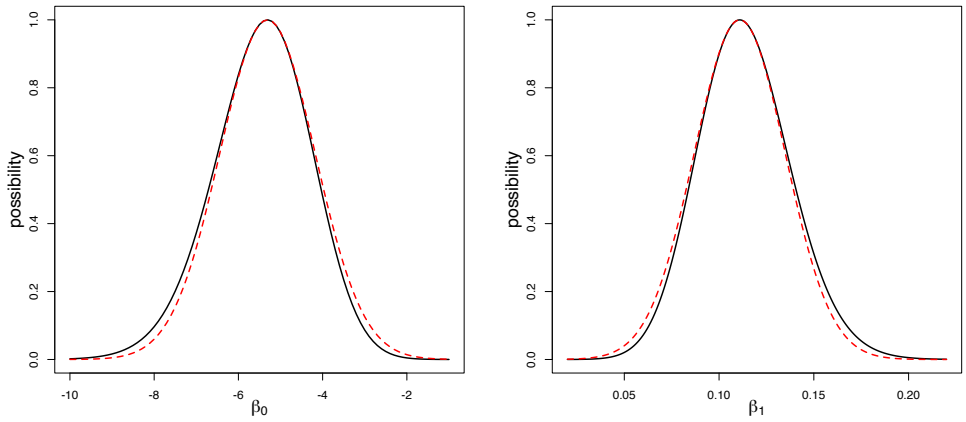


Figure 2: Contours are levels 0.1, 0.3, 0.5, 0.7 and 0.9 of the possibility distribution $\widetilde{\beta}$ for the `CHDAGE` dataset (black solid lines), and normal approximation (red broken lines).



(a) (b)

Figure 3: Marginal possibility distributions of β_0 and β_1 for the `CHDAGE` dataset (black solid lines), and normal approximations (red broken lines).

Regularization. Regularization is sometimes used with logistic regression to prevent overfitting [32]. For instance, L_2 regularized logistic regression is based on maximizing the penalized log-likelihood

$$\ell^{(R)}(\beta) = \ln L(\beta) - \frac{\xi}{2}\sum_{j=1}^{p}\beta_j^2, \tag{22}$$

where $\xi > 0$ is a hyperparameter. Denoting by $\widehat{\beta}^{(R)}$ the estimate of β found by maximizing (22), the possibility distribution of β can then be defined as

$$\pi_{\beta|\boldsymbol{y}}^{(R)}(\beta) = \frac{\exp[\ell^{(R)}(\beta;\boldsymbol{y})]}{\exp[\ell^{(R)}(\widehat{\beta}^{(R)};\boldsymbol{y})]} \propto \pi_{\beta|\boldsymbol{y}}(\beta)\pi_0(\beta), \tag{23}$$

with

$$\pi_0(\beta) = \exp\left(-\frac{1}{2}\beta^T\boldsymbol{\Xi}\beta\right),$$

where $\boldsymbol{\Xi}$ is the $(p+1)\times(p+1)$ diagonal matrix with diagonal terms $(0,\xi,\ldots,\xi)$. Function π_0 can be seen as a normal prior possibility distribution on β. Similary, L_1 regularization corresponds to the combination of the likelihood-based possibility distribution with a Laplace possibilistic prior. We note that this correspondance is purely formal: the penalization term in (22) usually does not encode true prior knowledge, and coefficient ξ is typically determined from the data using, e.g., cross-validation. Regularized logistic regression will no longer be mentioned in the rest of this paper. In Section 5, soft targets as proposed in [44] will be used to prevent overfitting by implicit regularization.

3.2. Estimation of posterior probabilities

Let us now assume that we observe a new feature vector $X = x$. The possibility distribution of the posterior probability $\theta(x)$ of $Y = 1$ given $X = x$ can be computed by applying the extension principle to (16) and (18); we get

$$\pi_{\theta(x)|\boldsymbol{y}}(\theta) = \sup_{\{\beta:\theta=[1+\exp(-\beta^T x)]^{-1}\}} \pi_{\beta|\boldsymbol{y}}(\beta). \tag{24}$$

Each value $\pi_{\theta(x)|\boldsymbol{y}}(\theta)$ can, thus, be found by maximizing (18) subject to the constraint $\theta = F_L(\beta^T x)$, which can be equivalently written as

$$\beta_0 = \mathsf{logit}(\theta) - \sum_{j=1}^{p} \beta_j x_j \tag{25}$$

with

$$\mathsf{logit}(\theta) = F_L^{-1}(\theta) = \ln \frac{\theta}{1-\theta}.$$

Substituting β_0 with the right-hand side of (25) in (18), we transform the constrained nonlinear optimization problem into an unconstrained one, which is the method proposed in [54].

Normal approximation. Alternatively, using the normal approximation (19) allows us to obtain an approximate closed-form expression for $\pi_{\theta(x)|\boldsymbol{y}}(\theta)$. Using Proposition 1 with (19) and $\boldsymbol{U} = x^T$, we can see that the possibility distribution of $z = x^T\beta$ is, approximately,

$$\pi_{z|\boldsymbol{y}} \sim \mathsf{GFN}\left(x^T\widehat{\beta}, (x^T[\boldsymbol{\mathcal{I}}(\widehat{\beta})]^{-1}x)^{-1}\right).$$

The possibility distribution of $\theta(x)$ can, thus, be approximated using the extension principle by

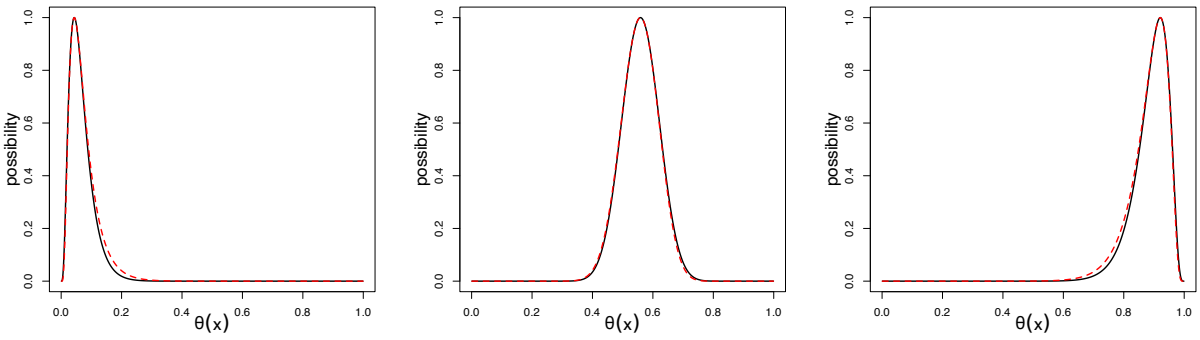
$$\begin{aligned}\pi_{\theta(x)|\boldsymbol{y}}(\theta) &= \sup_{\{z\in\mathbb{R}:\theta=[1+\exp(-z)]^{-1}\}} \pi_{z|\boldsymbol{y}}(z) \\ &= \pi_{z|\boldsymbol{y}}(\mathsf{logit}(\theta)) \\ &\approx \exp\left(-\frac{1}{2}(x^T[\boldsymbol{\mathcal{I}}(\widehat{\beta})]^{-1}x)^{-1}(\mathsf{logit}(\theta) - x^T\widehat{\beta})^2\right).\end{aligned} \tag{26}$$

Furthermore, from Proposition 1, the value of β maximizing (19) subject to $x^T\beta = \mathsf{logit}(\theta)$ is

$$\beta^* = \widehat{\beta} + [\boldsymbol{\mathcal{I}}(\widehat{\beta})]^{-1}x(x^T[\boldsymbol{\mathcal{I}}(\widehat{\beta})]^{-1}x)^{-1}(\mathsf{logit}(\theta) - x^T\widehat{\beta}).$$

This value can be used as a starting point in the optimization to compute the exact value of $\pi_{\theta(x)|\boldsymbol{y}}(\theta)$.

Example 3. *Continuing Example 2, Figure 4 shows the exact and approximate possibility distributions of $\theta(\mathsf{age})$ with $\mathsf{age} \in \{20, 50, 70\}$ for the* `CHDAGE` *dataset. The normal approximation is quite accurate, specially when the MLE $\widehat{\theta}(x)$ is not too close to 0 or 1.*



(a) (b) (c)

Figure 4: Possibility distributions of $\theta(\mathsf{age})$ with $\mathsf{age} = 20$ (a), $\mathsf{age} = 50$ (b) and $\mathsf{age} = 70$ (c) for the `CHDAGE` dataset (black solid lines), and normal approximations (red broken lines).

Table 1: Predictive mass functions for three values of age (Example 4).

	Exact			Approximate		
age	$m_{\widetilde{Y}(x)}(\{0\})$	$m_{\widetilde{Y}(x)}(\{1\})$	$m_{\widetilde{Y}(x)}(\{0,1\})$	$m_{\widetilde{Y}(x)}(\{0\})$	$m_{\widetilde{Y}(x)}(\{1\})$	$m_{\widetilde{Y}(x)}(\{0,1\})$
20	0.904	0.0201	0.0764	0.898	0.0209	0.0814
50	0.363	0.481	0.155	0.366	0.480	0.155
70	0.0393	0.841	0.120	0.0405	0.835	0.125

3.3. Prediction

Let us now consider the problem of quantifying the uncertainty on the response Y for a given x. As this response is random, this is a *prediction* problem. Given $X = x$, Y has a Bernoulli distribution $B(\theta(x))$: consequently, the expressions derived in Example 1 are still valid, replacing θ by $\theta(x)$. From (15), the expression of the predictive mass function is

$$m_{\widetilde{Y}(x)}(\{0\}) = 1 - \widehat{\theta}(x) - \int_{\widehat{\theta}(x)}^{1} \pi_{\theta(x)|\boldsymbol{y}}(u)du \tag{27a}$$

$$m_{\widetilde{Y}(x)}(\{1\}) = \widehat{\theta}(x) - \int_{0}^{\widehat{\theta}(x)} \pi_{\theta(x)|\boldsymbol{y}}(u)du \tag{27b}$$

$$m_{\widetilde{Y}(x)}(\{0,1\}) = \int_{0}^{1} \pi_{\theta(x)|\boldsymbol{y}}(u)du. \tag{27c}$$

These expressions were already obtained in [54]. The integrals in (27) can be computed by numerical integration, using either the exact possibility distribution $\pi_{\theta(x)|\boldsymbol{y}}$, or its approximate expression (26).

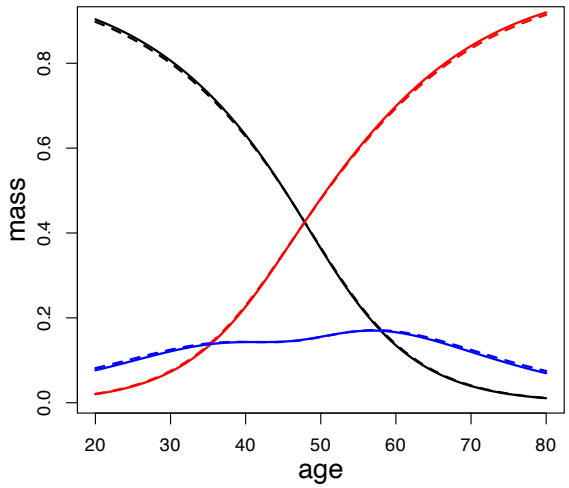


Figure 5: Predictive mass functions plotted against age for the CHDAGE dataset. Masses assigned to $\{0\}$, $\{1\}$ and $\{0,1\}$ are plotted, respectively, in black, red and blue. The exact and approximate values are plotted, respectively, as solid and broken lines.

Example 4. *Table 1 shows exact and approximate values of predictive mass functions $m_{\widetilde{Y}(x)}$ for the* CHDAGE *dataset and* age $\in \{20, 50, 70\}$. *The masses are plotted against* age *in Figure 5. We can remark that the normal approximations are quite accurate.*

4. Multinomial logistic regression

In this section, we consider the extension to multinomial logistic regression of the methodology described in Section 3. The estimation of coefficients and posterior probabilities will first be addressed, respectively, in Sections 4.1 and 4.2. The prediction problem will then be tackled in Section 4.3.

4.1. Estimation of coefficients

We consider a classification problem in which the response Y is a categorical variable taking values in $\mathcal{Y} = \{1, \ldots, K\}$ with $K \geq 3$. Let Ψ_S denote the *softmax transformation* from $\mathbb{R}^{K-1}$ to the simplex $\mathcal{S}_K$ of K-dimensional probability vectors, defined as

$$\psi_S(z_2, \ldots, z_K) = \left[\frac{1}{1 + \sum_{k=2}^{K} \exp(z_k)}, \frac{\exp(z_2)}{1 + \sum_{k=1}^{K} \exp(z_k)}, \ldots, \frac{\exp(z_K)}{1 + \sum_{k=2}^{K} \exp(z_k)} \right]^T.$$

Let $\theta_k(x; \beta)$ denote the conditional probability that $Y = k$ given $X = x$, where, as before, X denotes the extended feature vector. The vector of conditional probabilities is assumed to be given by

$$(\theta_1(x; \beta), \ldots, \theta_K(x; \beta)) = \psi_S(\beta_2^T x, \ldots, \beta_K^T x),$$

where $\beta_k \in \mathbb{R}^{p+1}$ is a vector of coefficients specific to class k, and $\beta = (\beta_2^T, \ldots, \beta_K^T)^T \in \mathbb{R}^{(K-1)(p+1)}$ is the vector of all parameters in the model. The conditional likelihood can then be written as

$$L(\beta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \theta_k(x_i; \beta)^{y_{ik}}, \tag{28}$$

where $y_{ik} = 1$ if $y_i = k$, and $y_{ik} = 0$ otherwise. As in the binomial case, the MLE $\widehat{\beta}$ of β can be found by maximizing (28) using an iterative nonlinear optimization algorithm, and the possibility distribution $\pi_{\beta|\boldsymbol{y}}$ of β is given by (18).

The observed information matrix is now [33, page 272]:

$$\boldsymbol{\mathcal{I}}(\widehat{\beta}) = \underset{\sim}{\boldsymbol{X}}^T \underset{\sim}{\boldsymbol{V}} \underset{\sim}{\boldsymbol{X}},$$

with $\underset{\sim}{\boldsymbol{X}}$ the $n(K-1) \times (p+1)(K-1)$ block matrix

$$\underset{\sim}{\boldsymbol{X}} = \mathbf{Id}_{K-1} \otimes \boldsymbol{X},$$

where $\mathbf{Id}_{K-1}$ is the identity matrix of size $(K-1) \times (K-1)$, $\otimes$ denotes the Kronecker product, and $\underset{\sim}{\boldsymbol{V}}$ is the $n(K-1) \times n(K-1)$ block matrix

$$\underset{\sim}{\boldsymbol{V}} = \begin{pmatrix} \boldsymbol{V}_{1,1} & \boldsymbol{V}_{1,2} & \cdots & \boldsymbol{V}_{1,(K-1)} \\ \boldsymbol{V}_{2,1} & \boldsymbol{V}_{2,2} & \cdots & \boldsymbol{V}_{2,(K-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{V}_{(K-1),1} & \boldsymbol{V}_{(K-1),2} & \cdots & \boldsymbol{V}_{(K-1),(K-1)} \end{pmatrix},$$

where each submatrix $\boldsymbol{V}_{kl}$ is an $n \times n$ diagonal matrix, the i-th diagonal element of $\boldsymbol{V}_{kk}$ is $\widehat{\theta}_{ki}(1 - \widehat{\theta}_{ki})$ with $\widehat{\theta}_{ki} = \theta_k(x_i, \widehat{\beta})$ and the i-th diagonal element of $\boldsymbol{V}_{kl}$ with $k \neq l$ is $-\widehat{\theta}_{ki}\widehat{\theta}_{li}$. The normal approximation formulas given in Section 3.1, in particular (19) for the possibility distribution of β and (21) for marginal possibility distributions, are unchanged.

4.2. Estimation of posterior probabilities

As before, we will first consider the estimation of conditional class probabilities for a given feature vector x, before addressing the problem of predicting the multinomial response Y in Section 4.3.

For a test vector x, let $\theta(x) = (\theta_1(x), \ldots, \theta_K(x))$ denote the vector of conditional class probabilities. The possibility distribution $\pi_{\theta(x)|\boldsymbol{y}}$ can be obtained from the possibility distribution $\pi_{\beta|\boldsymbol{y}}$ of β using the extension principle,

$$\pi_{\theta(x)|\boldsymbol{y}}(\theta) = \sup_{\{\beta: \theta = \psi_S(x^T\beta_2, \ldots, x^T\beta_K)\}} \pi_{\beta|\boldsymbol{y}}(\beta) \tag{29a}$$

$$= \sup_{\{\beta: \psi_S^{-1}(\theta) = (x^T\beta_2, \ldots, x^T\beta_K)\}} \pi_{\beta|\boldsymbol{y}}(\beta), \tag{29b}$$

where the inverse of the softmax transformation is

$$\psi_S^{-1}(\theta) = \left[\ln \frac{\theta_2}{\theta_1}, \ldots, \ln \frac{\theta_K}{\theta_1}\right]^T.$$

The possibility degree $\pi_{\theta(x)|\boldsymbol{y}}(\theta)$ can, thus, be computed by maximizing $\pi_{\beta|\boldsymbol{y}}(\beta)$ subject to $K-1$ linear equality constraints

$$x^T\beta_k = \ln\frac{\theta_k}{\theta_1}, \quad k = 2,\ldots,K. \tag{30}$$

From each constraint (30), we get

$$\beta_{k0} = \ln\left(\frac{\theta_k}{\theta_1}\right) - \sum_{j=1}^{p} x_j\beta_{kj}. \tag{31}$$

Replacing each β_{k0} for $k = 2,\ldots,K$ by the right-hand side of (31) in the expression of $\pi_{\beta|\boldsymbol{y}}(\beta)$ transforms the constrained optimization problem into an unconstrained one.

Normal approximation. An approximate closed-form expression for $\pi_{\theta(x)|\boldsymbol{y}}(\theta)$ can be obtained from the normal approximation (19) of $\pi_{\beta|\boldsymbol{y}}$ as follows. Let $z = (x^T\beta_2,\ldots,x^T\beta_K)$. We can write $z = \boldsymbol{U}_x\beta$, where $\boldsymbol{U}_x$ is matrix of size $(K-1)\times(K-1)(p+1)$ defined as

$$\boldsymbol{U}_x = \mathbf{Id}_{K-1}\otimes x^t.$$

Applying Proposition 1 to (19) with $\boldsymbol{U} = \boldsymbol{U}_x$, we obtain the approximate possibility distribution of z as

$$\pi_{z|\boldsymbol{y}} \sim \mathsf{GFV}\left(\boldsymbol{U}_x\widehat{\beta}, (\boldsymbol{U}_x[\boldsymbol{\mathcal{I}}(\widehat{\beta})]^{-1}\boldsymbol{U}_x^T)^{-1}\right).$$

The possibility distribution of $\theta(x)$ can, thus, be approximated by

$$\begin{aligned}\pi_{\theta(x)|\boldsymbol{y}}(\theta) &= \sup_{\{z:\theta=\psi_S(z)\}} \pi_{z|\boldsymbol{y}}(z)\\ &= \pi_{z|\boldsymbol{y}}(\psi_S^{-1}(\theta))\\ &\approx \exp\left(-\frac{1}{2}(\psi_S^{-1}(\theta)-\boldsymbol{U}_x\widehat{\beta})^T(\boldsymbol{U}_x[\boldsymbol{\mathcal{I}}(\widehat{\beta})]^{-1}\boldsymbol{U}_x^T)^{-1}(\psi_S^{-1}(\theta)-\boldsymbol{U}_x\widehat{\beta})\right).\end{aligned} \tag{32}$$

From Proposition 1, the corresponding value of β is

$$\beta^* = \widehat{\beta} + [\boldsymbol{\mathcal{I}}(\widehat{\beta})]^{-1}\boldsymbol{U}_x^T(\boldsymbol{U}_x[\boldsymbol{\mathcal{I}}(\widehat{\beta})]^{-1}\boldsymbol{U}_x^T)^{-1}(\psi_S^{-1}(\theta)-\boldsymbol{U}_x\widehat{\beta}).$$

This value can be used as a starting point to solve the constrained optimization problem (29).

Example 5. *The* `Wine` *dataset [25] contains the results of a chemical analysis of* $n = 178$ *wines grown in the same region in Italy but derived from three different cultivars. Here, we used only the first two features to allow easy display of the data in feature space (see Figure 6). Figure 7 shows contours of the exact and approximate possibility distributions* $\pi_{\theta(x)|\boldsymbol{y}}$ *in the three-dimensional probability simplex for two different values of* x*. As we can see, the normal approximation* (32) *is quite accurate.*

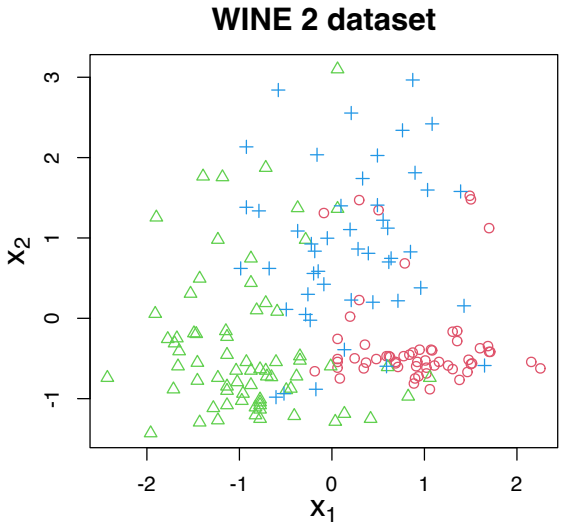
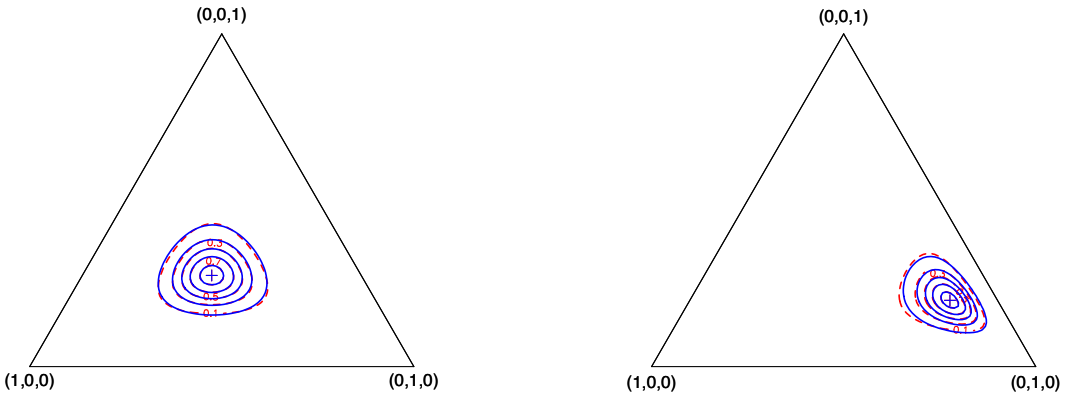
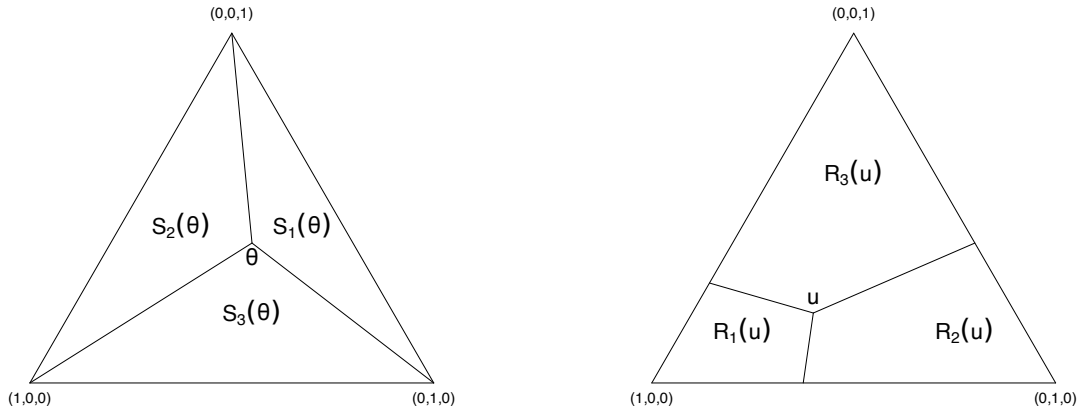


Figure 6: The `Wine` dataset in the space defined by the first two features.



(a) (b)

Figure 7: Contours at levels 0.1, 0.3, 0.5, 0.7 and 0.9 of possibility distributions $\pi_{\theta(x)|\boldsymbol{y}}$ for $x = (0.1, -0.5)$ (a) and $x = (-0.35, -0.47)$ (b) for the `Wine` dataset (blues solid lines), and normal approximations (red broken lines).



(a) (b)

Figure 8: Regions $S_k(\theta)$ (a) and $R_k(u)$ (b) in the three-dimensional simplex.

4.3. Prediction

φ-equation. To quantify the uncertainty on the random response Y for a given x, we need a φ-equation similar to (12). Such an equation was proposed by Dempster in [6] (see also [35] for a more recent description). Let U be a random vector having a uniform distribution in the probability simplex $\mathcal{S}_K$. Denote by $v_1, \ldots, v_K$ the K vertices of $\mathcal{S}_K$ such that $v_1 = (1, 0, \ldots, 0)$, $v_2 = (0, 1, 0, \ldots, 0)$, $\ldots$, $v_K = (0, \ldots, 0, 1)$. For each $\theta \in \mathcal{S}_K$, let $S_k(\theta)$ be the "subsimplex" obtained as the polytope with the same vertices as $\mathcal{S}_K$ except that vertex v_k is replaced by θ (see Figure 8a). It can be checked that $P(U \in S_k) = \theta_k$. Consequently, the random variable Y^* defined by

$$Y^* = \varphi(\theta, U) = \sum_{k=1}^{K} k \cdot I(U \in S_k(\theta)), \tag{33}$$

where $I(\cdot)$ denotes the indicator function, has the same distribution as Y. Regions $S_k(\theta)$ can be characterized by the following proposition [6].

Proposition 2. *For any $k \in \mathcal{Y}$, $\theta \in \mathcal{S}_K$ and $u \in \mathcal{S}_K$,*

$$u \in S_k(\theta) \Leftrightarrow \forall l \in \mathcal{Y}, \quad \frac{u_l}{u_k} \geq \frac{\theta_l}{\theta_k}. \tag{34}$$

Conversely, for any $u \in \mathcal{S}_K$ and $k \in \mathcal{Y}$, we denote by $R_k(u)$ the set of probability vectors θ such that $u \in S_k(\theta)$ (see Figure 8b). The regions $R_k(u)$ are also characterized by (34), where u is now held constant and θ varies.

Predictive belief functions. For fixed $U = u$, (33) allows us to compute the possibility distribution $\pi_{Y|\boldsymbol{y},x,u}$ on Y from the possibility distribution $\pi_{\theta(x)|\boldsymbol{y}}$ on $\theta(x)$ using the extension principle, as

$$\pi_{Y|\boldsymbol{y},x,u}(y) = \sup_{\{\theta\in\Theta:\varphi(\theta,u)=y\}} \pi_{\theta(x)|\boldsymbol{y}}(\theta) = \sup_{\theta\in R_y(u)} \pi_{\theta(x)|\boldsymbol{y}}(\theta) \tag{35}$$

for any $y \in \mathcal{Y}$. Taking into account the randomness of U, the mapping

$$\begin{aligned} \widetilde{Y}(x): \quad [0,1] &\to [0,1]^{\mathcal{Y}} \\ u &\mapsto \pi_{Y|\boldsymbol{y},x,u} \end{aligned}$$

is a RFS. The corresponding predictive plausibility and belief functions are defined, respectively, as

$$Pl_{\widetilde{Y}(x)}(A) = \mathbb{E}_U\left[\max_{y\in A} \pi_{Y|\boldsymbol{y},x,U}(y)\right] \quad \text{and} \quad Bel_{\widetilde{Y}(x)}(A) = 1 - \mathbb{E}_U\left[\max_{y\notin A} \pi_{Y|\boldsymbol{y},x,U}(y)\right]$$

for all $A \subseteq \mathcal{Y}$. These functions can be approximated by Monte Carlo simulation, drawing N independent realizations $u_1, \ldots, u_N$ of U and computing the possibility distributions $\pi_{Y|\boldsymbol{y},x,u_i}$ for each i. We then have

$$Pl_{\widetilde{Y}(x)}(A) \approx \frac{1}{N}\sum_{i=1}^{N} \max_{y\in A} \pi_{Y|\boldsymbol{y},x,u_i}(y), \quad Bel_{\widetilde{Y}(x)}(A) \approx 1 - \frac{1}{N}\sum_{i=1}^{N} \max_{y\notin A} \pi_{Y|\boldsymbol{y},x,u_i}(y). \tag{36}$$

We note that, for large K, we may not need to compute the whole belief and plausibility functions, depending on the chosen decision rule [15]. For instance, in many cases, it may be sufficient to compute the belief or the plausibility of singletons. (See Section 5.1 below).

Exact computation of $\pi_{Y|\boldsymbol{y},x,u}(y)$. To compute $\pi_{Y|\boldsymbol{y},x,u}(y)$, we can proceed as follows. From (29), $\pi_{\theta(x)|\boldsymbol{y}}(\theta)$ can be computed by maximizing $\pi_{\beta|\boldsymbol{y}}(\beta)$ subject to

$$x^T\beta_k = \ln\frac{\theta_k}{\theta_1}, \quad k = 2, \ldots, K.$$

Now, from Proposition 2, for $k > 1$,

$$\begin{aligned} \theta \in R_k(u) &\Leftrightarrow \forall l \in \mathcal{Y}, \quad \frac{\theta_l}{\theta_k} \leq \frac{u_l}{u_k} \\ &\Leftrightarrow \forall l \in \mathcal{Y}, \quad \ln\frac{\theta_l}{\theta_1} - \ln\frac{\theta_k}{\theta_1} \leq \ln\frac{u_l}{u_k}. \end{aligned}$$

Consequently, $\pi_{Y|\boldsymbol{y},x,u}(1)$ can be computed by solving the nonlinear optimization problem

$$\max_{\beta} \pi_{\beta|\boldsymbol{y}}(\beta) \tag{37a}$$

subject to

$$x^T\beta_l \leq \ln\frac{u_l}{u_1}, \quad l = 2, \ldots, K, \tag{37b}$$

and $\pi_{Y|\boldsymbol{y},x,u}(k)$ for $k > 1$ can be computed by maximizing (37a) subject to

$$-x^T\beta_k \leq \ln\frac{u_1}{u_k} \tag{37c}$$

$$x^T(\beta_l - \beta_k) \leq \ln\frac{u_l}{u_k}, \quad l \in \{2,\ldots,K\}\setminus\{k\}. \tag{37d}$$

We note that, trivially, $\pi_{Y|\boldsymbol{y},x,u}(k) = 1$ iff $\widehat{\theta}(x) = \psi_S(x^T\widehat{\beta}_2,\ldots,x^T\widehat{\beta}_K) \in R_k$. We, thus, need to solve only $K-1$ constrained optimization problems.

Approximate computation of $\pi_{Y|\boldsymbol{y},x,u}(y)$. Using the normal approximation (32), an approximation of $\pi_{Y|\boldsymbol{y},x,u}(k)$ can be obtained by minimizing the quadratic function

$$(z - \boldsymbol{U}_x\widehat{\beta})^T(\boldsymbol{U}_x[\boldsymbol{\mathcal{I}}(\widehat{\beta})]^{-1}\boldsymbol{U}_x^T)^{-1}(z - \boldsymbol{U}_x\widehat{\beta}) \tag{38a}$$

subject to linear constraints

$$z_l \leq \ln\frac{u_l}{u_1}, \quad l = 2,\ldots,K, \tag{38b}$$

for $k = 1$ and

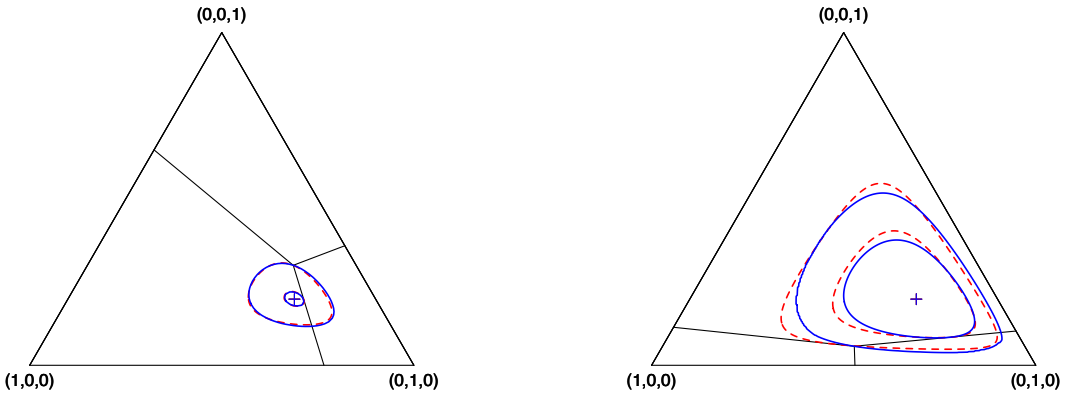
$$-z_k \leq \ln\frac{u_1}{u_k} \tag{38c}$$

$$z_l - z_k \leq \ln\frac{u_l}{u_k}, \quad l \in \{2,\ldots,K\}\setminus\{k\} \tag{38d}$$

for $k > 1$. These are quadratic optimization problems, which can be solved very efficiently. Furthermore, the dimension of z is $K-1$, whereas the dimension of β is $(p+1)(K-1)$.

Example 6. *Figure 9 illustrates the result of the optimization for the* `Wine` *dataset, with* $x = (-0.186, -0.659)$ *and* $\widehat{\theta}(x) = (0.211, 0.590, 0.199)$. *In Figure 9a,* $u = (0.164, 0.536, 0.300)$ *and* $\widehat{\theta}(x) \in R_1(u)$. *Consequently,* $\pi_{Y|\boldsymbol{y},x,u}(1) = 1$. *The inner and outer blue solid curves are contours of* $\pi_{\theta(x)|\boldsymbol{y}}$ *corresponding to the solutions of Problem* (37) *for, respectively,* $k = 2$ *and* $k = 3$. *We get* $\pi_{Y|\boldsymbol{y},x,u}(2) = 0.929$ *and* $\pi_{Y|\boldsymbol{y},x,u}(3) = 0.245$. *The broken red curves are the contours of the approximation of* $\pi_{\theta(x)|\boldsymbol{y}}$ *corresponding to the solutions of quadratic optimization problem* (38)*; we get the approximations* $\pi_{Y|\boldsymbol{y},x,u}(2) \approx 0.928$ *and* $\pi_{Y|\boldsymbol{y},x,u}(3) \approx 0.264$. *Figure 9b corresponds to* $u = (0.444, 0.499, 0.057)$. *As* $\widehat{\theta}(x) \in R_3(u_i)$, $\pi_{Y|\boldsymbol{y},x,u}(3) = 1$. *The solution of Problem* (37) *gives us* $\pi_{Y|\boldsymbol{y},x,u}(1) = 5.48\times10^{-5}$ *and* $\pi_{Y|\boldsymbol{y},x,u}(2) = 3.13\times10^{-2}$*; the approximate solutions are* $\pi_{Y|\boldsymbol{y},x,u}(1) \approx 5.35\times10^{-5}$ *and* $\pi_{Y|\boldsymbol{y},x,u}(2) \approx 1.88\times10^{-2}$.

Table 2 reports the exact and approximate predictive plausibility, belief and mass functions computed using (36) *with* $N = 5000$ *and the solutions of problems* (37) *or* (38). *Figure 10 displays a scatter plot of exact versus approximate mean predicted masses for 50 vectors* x_i *randomly chosen from the* `Wine` *dataset, confirming the very good quality of the approximation. Figures 11 and 12 show, respectively, contour plots of the plausibility of the singletons and the masses assigned to each of the seven nonempty focal sets, for the approximate predictive belief functions.*



(a) (b)

Figure 9: Solution of constrained optimization problems for the calculation of $\pi_{Y|\boldsymbol{y},x,u}(k)$, for two different values of u (see details in text). The solid blue curves are exact contours of possibility distributions $\pi_{\theta(x)|\boldsymbol{y}}$, while the broken red curves are based on the normal approximation of $\pi_{\beta|\boldsymbol{y}}(\beta)$.

Table 2: Exact and approximate predictive plausibility, belief and mass functions for the `Wine` dataset with $x = (-0.186, -0.659)$ (Example 6).

	A	{1}	{2}	{1,2}	{3}	{1,3}	{2,3}	{1,2,3}
Exact	$Pl_{\widetilde{Y}(x)}(A)$	0.313	0.701	0.867	0.292	0.5359	0.8580	1
	$Bel_{\widetilde{Y}(x)}(A)$	0.142	0.464	0.708	0.133	0.2993	0.6865	1
	$m_{\widetilde{Y}(x)}(A)$	0.142	0.464	0.102	0.133	0.0244	0.0895	0.0449
Approx.	$Pl_{\widetilde{Y}(x)}(A)$	0.317	0.697	0.866	0.293	0.5398	0.8557	1
	$Bel_{\widetilde{Y}(x)}(A)$	0.144	0.460	0.707	0.134	0.3033	0.6826	1
	$m_{\widetilde{Y}(x)}(A)$	0.144	0.460	0.102	0.134	0.0249	0.0883	0.0462

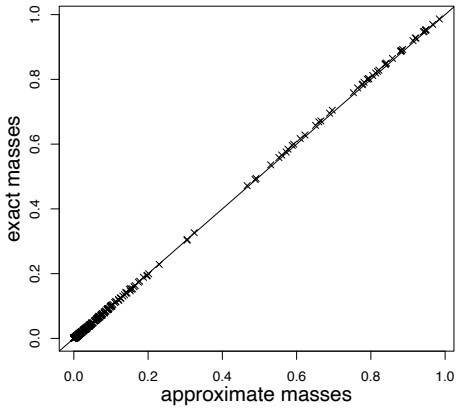


Figure 10: Exact versus approximate predicted masses for 50 instances randomly chosen from the `Wine` dataset.

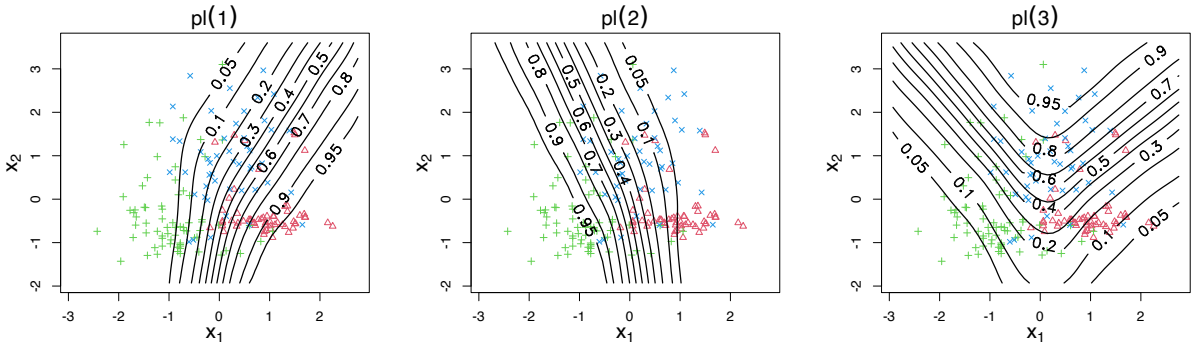


Figure 11: Contour plots of the plausibility of singletons for the `Wine` dataset.

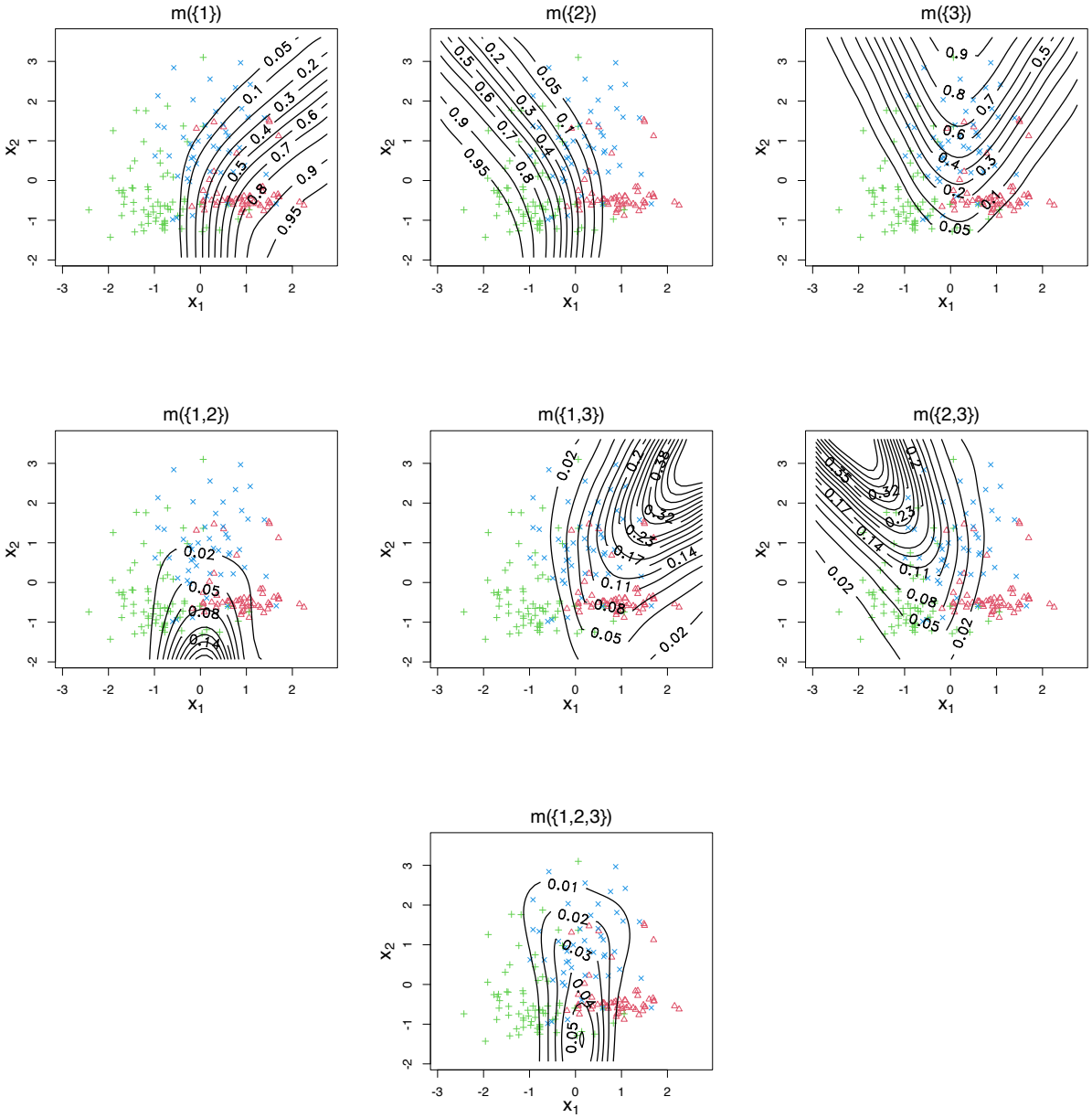


Figure 12: Contour plots of the masses assigned to the seven nonempty subsets of $\{1, 2, 3\}$ for the `Wine` dataset.

5. Numerical experiments

In this section, we describe some experiments aiming to determine whether the predictive belief functions obtained by evidential logistic regression, as explained in Sections 3 and 4, are more informative, in some sense, than the probabilistic outputs of conventional logistic regression. This could be done using, at least, two approaches. The first one is to combine the outputs of evidential and probabilistic classifiers with those of other classifiers and compare the accuracies of the predictions; this approach was used in [54] and [42] to demonstrate the advantages of evidential calibration of single or multiple binary SVM classifiers, as compared to other calibration methods. Another approach, which will be adopted here, is to consider decision rules with rejection, and to compare error rates for various rejection rates obtained with predictive belief functions on the one hand, and estimated posterior probabilities on the other hand. As a given rejection rate is achieved by comparing the maximum degree belief or plausibility to some threshold, the error-reject curve, by considering all possible thresholds, characterizes the "information content" of the predictive belief function better than the error rate without rejection alone[3]. The datasets and the experimental settings will first be described in Section 5.1. The results will then be presented and discussed in Section 5.2. Finally, a comparison with the evidence-based predictive belief functions introduced in [16] is sketched in Section 5.3.

5.1. Data and experimental settings

Datasets. We considered two simulated datasets and six real datasets as summarized in Table 3. The Pima Indians Diabetes dataset can be retrieved from Kaggle[4]. The other real datasets are available from the UCI Machine Learning repository[5]. The dimensions of the iono and sonar datasets were reduced, respectively, to 15 and 20 by principal component analysis. For the glass dataset, we considered only the three classes with the largest numbers of observations. For the vowel and letter datasets, we considered only the first six classes to reduce computation time.

The two-class simulated data were composed of 1000 observations generated from 10-dimensional normal distributions with means $\mu_1 = (0, \ldots, 0)^T$, $\mu_2 = (1, \ldots, 1)^T$, covariance matrices $\mathbf{\Sigma}_1 = \mathbf{Id}_{10}$, $\mathbf{\Sigma}_2 = 3\mathbf{Id}_{10}$, and equal prior probabilities. The simul4 data were generated from a mixture of four Gaussian distributions with means $\mu_1 = (0, \ldots, 0)^T$, $\mu_2 = (0.5, \ldots, 0.5)^T$, $\mu_3 = (0, 0.5, \ldots, 0, 0.5)^T$, $\mu_4 = (0.5, 0, , \ldots, 0.5, 0)^T$, covariance matrices $\mathbf{\Sigma}_1 = 0.5\mathbf{Id}_{20}$, $\mathbf{\Sigma}_2 = \mathbf{Id}_{20}$, $\mathbf{\Sigma}_3 = 2\mathbf{Id}_{20}$, $\mathbf{\Sigma}_4 = 3\mathbf{Id}_{20}$ and equal prior probabilities.

As the likelihood function becomes more and more concentrated around the true value of the parameter as the sample size increases, the difference between decisions made from estimated posterior probabilities on the one hand, and predictive belief functions on the other hand, is likely to be more pronounced in the case of small sample size. For each dataset, we thus randomly drew a small subsample and used the rest of the data for testing.

[3] A similar approach was used in [46] to evaluate the performance of evidential choquistic regression.

[4] `https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/`.

[5] `https://archive.ics.uci.edu`.

Table 3: Characteristics of the datasets.

Full name	Short name	# observations	# classes	# features
2-class simulated data	simul2	1000	2	10
Pima Indians Diabetes	pima	768	2	8
Ionosphere	ionosphere	351	2	33
Sonar, Mines vs. Rocks	sonar	208	2	60
4-class simulated data	simul4	300	4	20
Vowel recognition	vowel	540	6	10
Glass identification	glass	175	3	10
Letter identification	letter	600	6	16

The learning process was repeated $N = 50$ times with different random training sets, and the error-reject curves were averaged.

Decision rules. Another important issue is the choice of a decision rule. Decision rules for classification with rejection in the Dempster-Shafer setting were studied in [11] and the more general problem of partial classification was recently addressed in [38]. Here, the loss was assumed to be 0 for correct classification, 1 for misclassification, and $\lambda \in [0, 1]$ for rejection. By varying λ, different reject rates and associated error rates can be obtained and an error-reject curve can be plotted. Based on the predictive belief function, four decision rules with rejection were considered [11]:

1. The *pessimistic* rule minimizing the upper expected loss; according to this rule, an observation x is rejected if the maximum degree of belief is less than $1 - \lambda$, otherwise it is assigned to the class with the larger degree of belief;
2. The *optimistic* rule minimizing the lower expected loss: an observation x is rejected if the maximum plausibility is less than $1 - \lambda$, otherwise it is assigned to the class with the larger plausibility;
3. The *pignistic* rule [49], based on the pignistic probability distribution defined as

$$betp(k) = \sum_{\emptyset \neq A \subseteq \mathcal{Y}} I(k \in A) \frac{m(A)}{|A|}, \quad k = 1, \ldots, K.$$

   An observation x is assigned to the class with the maximum pignistic probability if $\max_k betp(k) \geq 1 - \lambda$, otherwise it is rejected;
4. The *normalized-plausibility* rule [4], based on the normalized plausibility

$$plp(k) = \frac{pl(k)}{\sum_{l=1}^{K} pl(l)}, \quad k = 1, \ldots, K.$$

   An observation x is assigned to the class with the maximum plausibility if $\max_k plp(k) \geq 1 - \lambda$, otherwise it is rejected.

The error-reject curves obtained with these evidential decision rules were compared to those obtained with the "plug-in probabilistic rule" based on estimated posterior probabilities $\widehat{\theta}(x)$.

Computation of predictive belief functions. The predictive belief functions were computed as explained in Section 3.3 and 4.3 using approximate possibility distributions of $\theta(x)$ (26) and (32). To avoid overfitting as well as numerical problems when maximizing the likelihood, the responses y_i were replaced by "soft targets" using Laplace smoothing, as proposed in [44]. Specifically, in the binomial case, y_i was replaced, in the expression of the conditional likelihood (17), by t_i defined as

$$t_i = \begin{cases} \frac{n_1+1}{n_1+2} & \text{if } y_i = 1, \\ \frac{1}{n_0+2} & \text{if } y_i = 0, \end{cases}$$

where n_0 and n_1 denote the number of observations from classes 0 and 1 in the training set. In the multinomial case, y_{ik} was replaced, in the expression of the conditional likelihood (28), by t_{ik} defined as

$$t_{ik} = \begin{cases} \frac{n_k+1}{n_k+K} & \text{if } y_i = k, \\ \frac{1}{n_k+K} & \text{otherwise,} \end{cases}$$

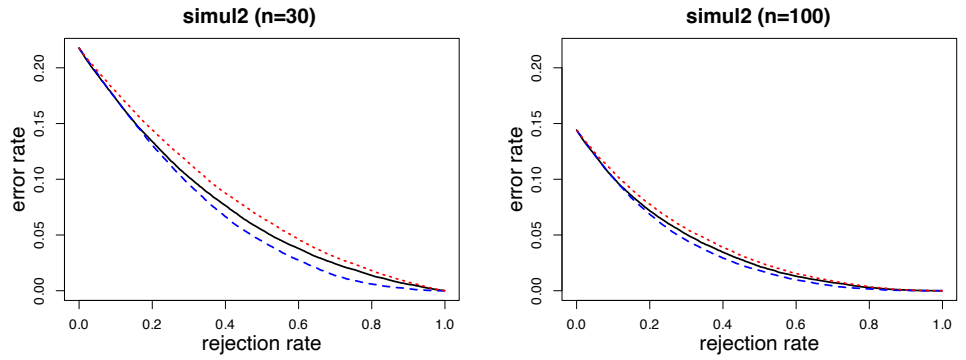
where n_k denotes the number of observations in class k.

5.2. Results

We hereafter discuss the results for binary and multi-class classification tasks separately.

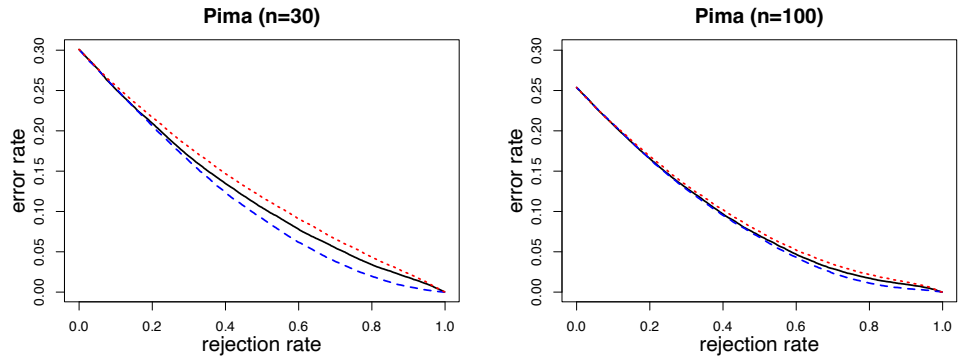
Binary classification. The error-reject curves for the binary classification tasks are shown in Figures 13 and 14. For each of the four two-class datasets, we plot the error-reject curves for plug-in probabilistic, pessimistic and optimistic decision rules, for training sets of sizes $n = 30$ and $n = 100$. The error-reject curves for the pignistic and normalized-plausibility rules are not shown, because they are almost exactly identical to those of the pessimistic rule. We can see that the pessimistic rule (as well as the pignistic and normalized-plausibility rules) perform better than the plug-in probabilistic rule. The difference is particularly large in the case of the ionosphere dataset (Figures 14a-14b), and it is more pronounced for $n = 30$ than for $n = 100$, as expected. In contrast, the optimistic decision rule performs worse than the plug-in probabilistic rule.

Multi-class classification. The error-reject curves for the multi-class classification tasks are shown in Figure 15. To avoid cluttering the graphs, we do not show the curves for the optimistic rule (which is always above the other curves) and for the normalized-plausibility rule, which always performs worse than the pignistic rule. For each of the vowel and letter datasets, we considered two learning tasks: classifying the first three classes (Figures 15c and 15e) and classifying all six classes (Figures 15d and 15f). We can see that the pessimistic rule has the best performance for the simul4 and glass datasets, as well as for the vowel and letter datasets with three classes. In contrast, the pignistic rule has the best performance for



(a)

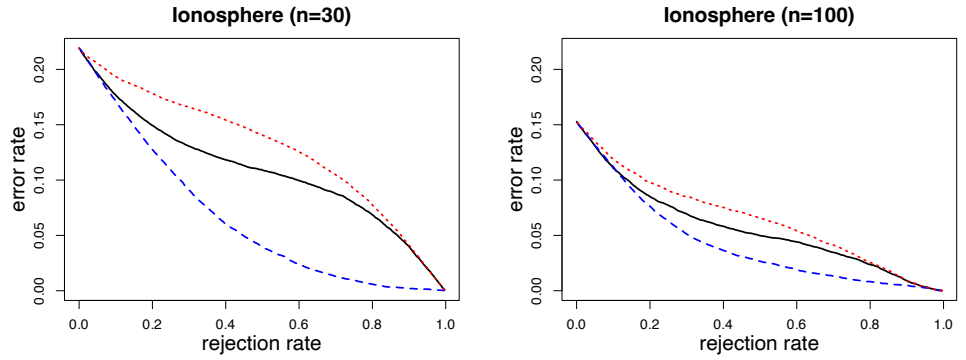
(b)



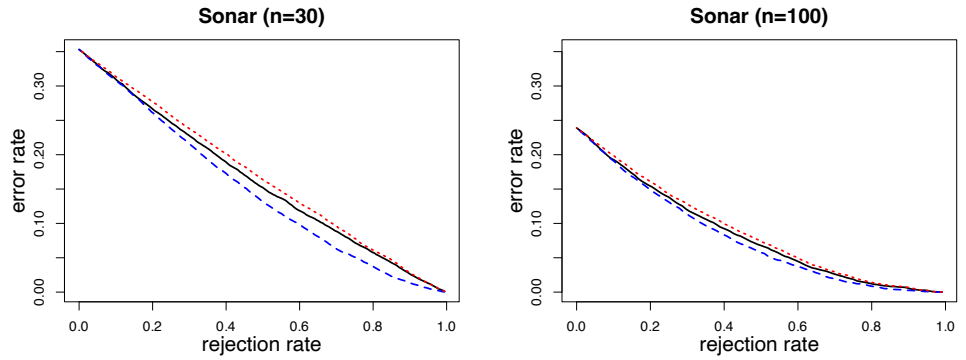
(c)

(d)

Figure 13: Error-reject curves for the simul2 and pima datasets, with $n = 30$ and $n = 100$ learning samples. The solid black, dashed blue and dotted red curves correspond, respectively, to the plug-in probabilistic, pessimistic and optimistic decision rules.



(a) (b)



(c) (d)

Figure 14: Error-reject curves for the ionosphere and sonar datasets, with $n = 30$ and $n = 100$ learning samples. The solid black, dashed blue and dotted red curves correspond, respectively, to the plug-in probabilistic, pessimistic and optimistic decision rules.

the vowel and letter datasets with six classes. The poor performance of the pessimistic rule with a larger number of classes can be explained by the fact that it uses only the masses assigned to the singletons, whereas the pignistic rule uses the whole mass functions.

All in all, these results show that the (approximate) predictive belief functions can be used to make decisions with lower rates for given rejection rates, as compared to the estimated posterior probabilities. This can be explained by the fact that the approximate predictive belief functions are computed using not only the estimated coefficients $\widehat{\beta}$, but also the observed information matrix $\boldsymbol{\mathcal{I}}(\widehat{\beta})$. The experimental results presented here show that the method to compute predictive belief functions introduced in Section 3 and 4 makes effective use of this additional information. In the next section, we briefly discuss the differences between the likelihood-based predictive belief functions studied in this paper and the latent belief functions introduced in [16], and we present some comparative results.

5.3. Comparison with latent belief functions

In [16], we showed that binomial and multinomial logistic regression can be interpreted as combining feature-based simple mass functions by Dempster's rule, resulting in a latent belief function, whose normalized plausibilities are the output posterior probabilities. The latent belief function has more degrees of freedom than the plug-in output probability distribution, but it is only based on the estimated coefficients $\widehat{\beta}$, as opposed to the likelihood-based belief function, which depends on the whole likelihood function (or only on the MLE and the observed information for the normal approximation). The latter belief function thus contains more information and can be expected to allow for more accurate decisions (at the cost of a significantly higher computational complexity). This hypothesis was confirmed experimentally for the datasets considered in this paper. For instance, Figure 16 shows the error-reject curves corresponding to the estimated posterior probabilities and to the latent belief functions with the pessimistic, optimistic and pignistic rules, for the simul2, ionosphere, simul4 and glass datasets. As can be seen from these results, the latent belief functions do not allow one to make better decisions than those based on the estimated posterior probabilities.

6. Conclusions

The likelihood-based approach to statistical inference introduced in [17, 19] treats the relative likelihood function as a possibility distribution in the parameter space. By expressing new data as a function of the parameter and a random variable with known probability distribution, one then defines a random fuzzy set and an associated predictive belief function representing uncertain knowledge about future observations. This approach yields the same results as Bayesian inference in the special case where prior knowledge about the parameter is given in the form of a probability distribution, but it does not require such prior information.

In this paper, we have applied this method of inference to binomial and multinomial logistic regression by first computing the possibility distribution of posterior class probabilities using Zadeh's extension principle. In the binomial case, degrees of belief and plausibility about the class of a new observation are then obtained by numerical integration of this possibility distribution, yielding the same solution as that described in [54]. In the multinomial

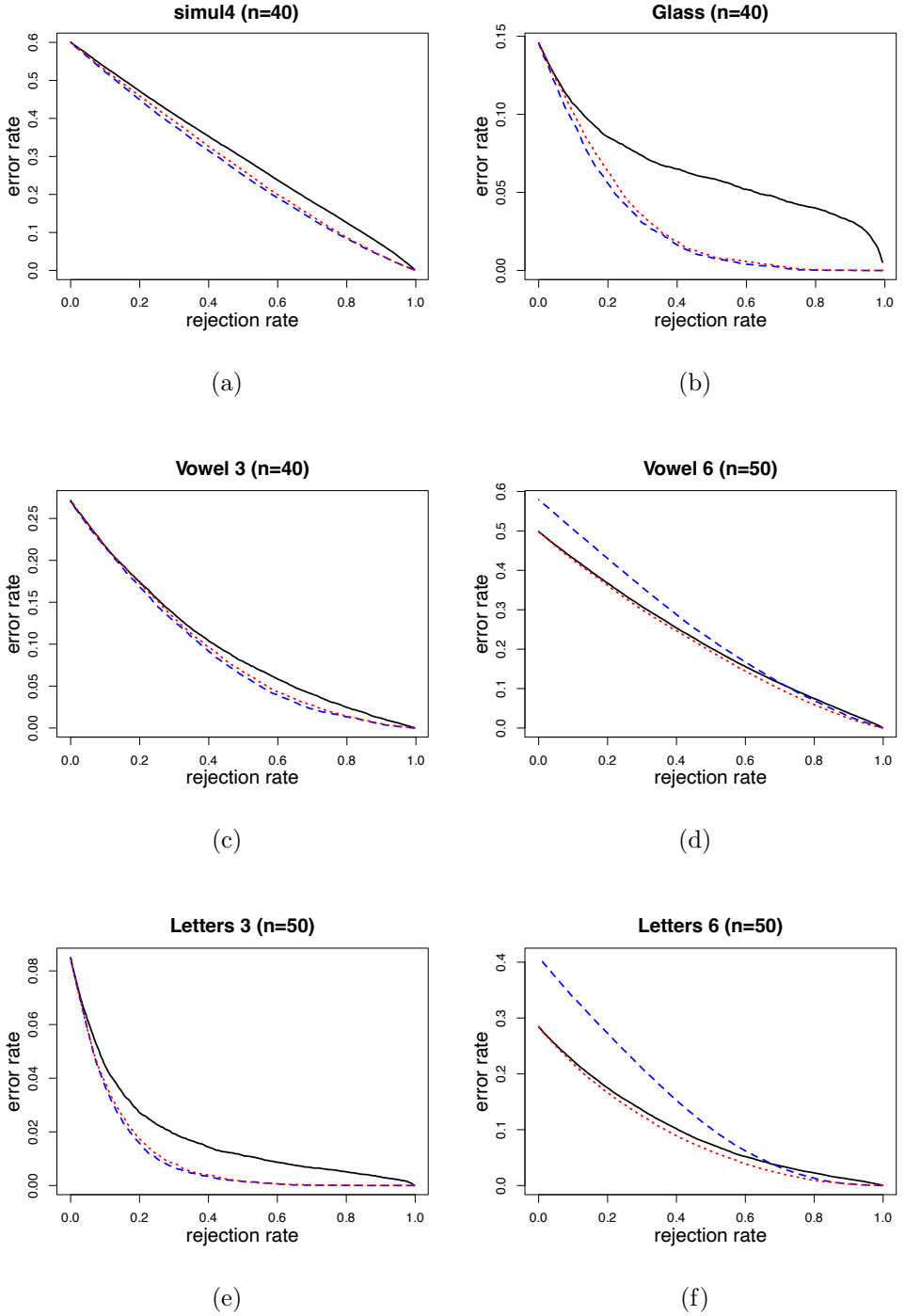


Figure 15: Error-reject curves for the simul4, glass, vowel (with three and six classes) and letter (with three and six classes) datasets. The solid black, dashed blue, dotted red curves correspond, respectively, to the plug-in probabilistic, pessimistic and pignistic decision rules.

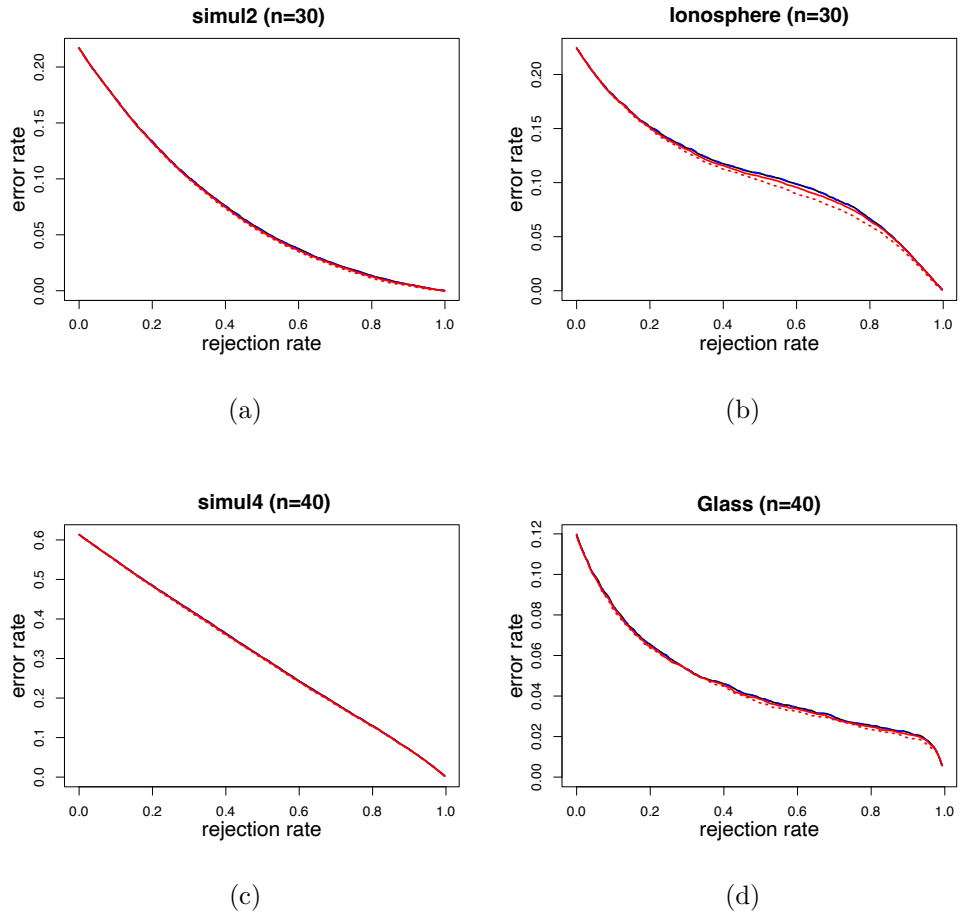


Figure 16: Error-reject curves for the simul2, ionosphere, simul4 and glass datasets. The solid black, dashed blue, dotted red and solid red curves correspond, respectively, to the plug-in probabilistic, pessimistic, optimistic and pignistic decision rules.

case, the calculations are more involved as they are based on a combination of nonlinear constrained optimization and Monte-Carlo simulation. In both cases, the computations can be considerably simplified using a normal approximation of the relative likelihood function, which has been shown to be usually quite accurate even for small sample sizes.

The predictive belief functions computed using this approach depend not only on the MLEs of the coefficients, but also on the whole likelihood function, or on the observed information if the normal approximation of the relative likelihood is used. Consequently, they are more informative than the plug-in posterior probability estimates. This was verified experimentally by showing that evidential decision rules based on these predictive belief functions achieve lower error rates for different reject rates, as compared to the decision rules based on estimated posterior probabilities or even the latent belief functions studied in [16].

Whereas logistic regression can only perform linear classification, it is at the basis of several nonlinear classification models such as kernel logistic regression [47], multi-layer neural networks [30] or choquistic regression [29]. An evidential version of binomial choquistic regression was already proposed in [46]. The method of inference demonstrated in this paper will be extended to other models in future work.

References

[1] V. Antoine, J. A. Guerrero, and J. Xie. Fast semi-supervised evidential clustering. *International Journal of Approximate Reasoning*, 133:116–132, 2021.

[2] L. Braglia. *aplore3: Datasets from Hosmer, Lemeshow and Sturdivant, "Applied Logistic Regression" (3rd Ed., 2013)*, 2016. R package version 0.9.

[3] L. Cella and R. Martin. Possibility-theoretic statistical inference offers performance and probativeness assurances. *International Journal of Approximate Reasoning*, 163:109060, 2023.

[4] B. R. Cobb and P. P. Shenoy. On the plausibility transformation method for translating belief function models to probability models. *International Journal of Approximate Reasoning*, 41(3):314–330, 2006.

[5] I. Couso and L. Sánchez. Upper and lower probabilities induced by a fuzzy random variable. *Fuzzy Sets and Systems*, 165(1):1–23, 2011.

[6] A. P. Dempster. New methods for reasoning towards posterior distributions based on sample data. *Annals of Mathematical Statistics*, 37:355–374, 1966.

[7] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[8] A. P. Dempster. A generalization of Bayesian inference (with discussion). *J. R. Statistical Society B*, 30:205–247, 1968.

[9] A. P. Dempster. The Dempster-Shafer calculus for statisticians. *International Journal of Approximate Reasoning*, 48(2):365–377, 2008.

[10] T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics*, 25(05):804–813, 1995.

[11] T. Denœux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition*, 30(7):1095–1107, 1997.

[12] T. Denœux. A neural network classifier based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics A*, 30(2):131–150, 2000.

[13] T. Denœux. Likelihood-based belief function: justification and some extensions to low-quality data. *International Journal of Approximate Reasoning*, 55(7):1535–1547, 2014.

[14] T. Denœux. 40 years of Dempster-Shafer theory. *International Journal of Approximate Reasoning*, 79:1–6, 2016.

[15] T. Denoeux. Decision-making with belief functions: a review. *International Journal of Approximate Reasoning*, 109:87–110, 2019.
[16] T. Denœux. Logistic regression, neural networks and Dempster-Shafer theory: A new perspective. *Knowledge-Based Systems*, 176:54–67, 2019.
[17] T. Denœux. Belief functions induced by random fuzzy sets: A general framework for representing uncertain and fuzzy evidence. *Fuzzy Sets and Systems*, 424:63–91, 2021.
[18] T. Denœux. NN-EVCLUS: neural network-based evidential clustering. *Information Sciences*, 572:297–330, 2021.
[19] T. Denœux. Reasoning with fuzzy and uncertain evidence using epistemic random fuzzy sets: General framework and practical models. *Fuzzy Sets and Systems*, 453:1–36, 2023.
[20] T. Denœux, D. Dubois, and H. Prade. Representations of uncertainty in artificial intelligence: Beyond probability and possibility. In P. Marquis, O. Papini, and H. Prade, editors, *A Guided Tour of Artificial Intelligence Research*, volume 1, chapter 4, pages 119–150. Springer Verlag, 2020.
[21] T. Denœux, D. Dubois, and H. Prade. Representations of uncertainty in artificial intelligence: Probability and possibility. In P. Marquis, O. Papini, and H. Prade, editors, *A Guided Tour of Artificial Intelligence Research*, volume 1, chapter 3, pages 69–117. Springer Verlag, 2020.
[22] T. Denœux, O. Kanjanatarakul, and S. Sriboonchitta. A new evidential k-nearest neighbor rule based on contextual discounting with partially supervised learning. *International Journal of Approximate Reasoning*, 113:287–302, 2019.
[23] T. Denœux and V. Kreinovich. Algebraic product is the only "and-like"-operation for which normalized intersection is associative: A proof. In *Fifth International Conference on Artificial Intelligence and Computational Intelligence (AICI 2024)*, Hanoi, Vietnam, January 2024. `https://hal.science/hal-04436177`.
[24] T. Denœux. Parametric families of continuous belief functions based on generalized gaussian random fuzzy numbers. *Fuzzy Sets and Systems*, 471:108679, 2023.
[25] D. Dua and C. Graff. UCI machine learning repository, 2017. `http://archive.ics.uci.edu/ml`.
[26] D. Dubois. Possibility theory and statistical reasoning. *Computational Statistics and Data Analysis*, 51(1):47–69, 2006.
[27] D. Dubois, H. T. Nguyen, and H. Prade. Possibility theory, probability and fuzzy sets: Misunderstandings, bridges and gaps. In D. Dubois and H. Prade, editors, *Fundamentals of Fuzzy sets*, pages 343–438. Kluwer Academic Publishers, Boston, 2000.
[28] D. Dubois and H. Prade. Possibility theory: qualitative and quantitative aspects. In D. M. Gabbay and P. Smets, editors, *Handbook of Defeasible reasoning and uncertainty management systems*, volume 1, pages 169–226. Kluwer Academic Publishers, Dordrecht, 1998.
[29] A. Fallah Tehrani, W. Cheng, K. Dembczyński, and E. Hüllermeier. Learning monotone nonlinear models using the Choquet integral. *Machine Learning*, 89(1):183–211, 2012.
[30] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.
[31] P. C. Groenewald and L. Mokgatlhe. Bayesian computation for logistic regression. *Computational Statistics & Data Analysis*, 48(4):857–868, 2005.
[32] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.
[33] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, 3rd edition, 2013.
[34] L. Huang, S. Ruan, P. Decazes, and T. Denœux. Lymphoma segmentation from 3D PET-CT images using a deep evidential network. *International Journal of Approximate Reasoning*, 149:39–60, 2022.
[35] P. E. Jacob, R. Gong, P. T. Edlefsen, and A. P. Dempster. A Gibbs sampler for a class of random convex polytopes. *Journal of the American Statistical Association*, 116(535):1181–1192, 2021.
[36] O. Kanjanatarakul, S. Sriboonchitta, and T. Denœux. Forecasting using belief functions: an application to marketing econometrics. *International Journal of Approximate Reasoning*, 55(5):1113–1128, 2014.
[37] O. Kanjanatarakul, S. Sriboonchitta, and T. Denœux. Prediction of future observations using belief functions: A likelihood-based approach. *International Journal of Approximate Reasoning*, 72:71–94,

2016.
[38] L. Ma and T. Denœux. Partial classification in the belief function framework. *Knowledge-Based Systems*, 214:106742, 2021.
[39] R. Martin. Valid and efficient imprecise-probabilistic inference with partial priors, I. first results, 2022. `https://doi.org/10.48550/arXiv.2203.06703`.
[40] R. Martin. Valid and efficient imprecise-probabilistic inference with partial priors, II. general framework, 2022. `https://doi.org/10.48550/arXiv.2211.14567`.
[41] R. Martin and C. Liu. *Inferential Models: Reasoning with Uncertainty*. CRC Press, Boca Raton, 2016.
[42] P. Minary, F. Pichon, D. Mercier, E. Lefevre, and B. Droit. Evidential joint calibration of binary SVM classifiers. *Soft Computing*, 23(13):4655–4671, 2019.
[43] K. B. Petersen and M. S. Pedersen. The matrix cookbook, November 2012. `http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html`.
[44] J. Platt. Probabilities for SV machines. In A. J. Smola, P. Bartlett, B. Schölkopf, and D. Schuurman, editors, *Advances in Large-Margin Classifiers*, pages 61–74. MIT Press, 2000.
[45] B. Quost, T. Denœux, and S. Li. Parametric classification with soft labels using the evidential EM algorithm: linear discriminant analysis versus logistic regression. *Advances in Data Analysis and Classification*, 11(4):659–690, Dec 2017.
[46] S. Ramel, F. Pichon, and F. Delmotte. A reliable version of choquistic regression based on evidence theory. *Knowledge-Based Systems*, 205:106252, 2020.
[47] B. Schölkopf and A. J. Smola. *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, Massachusetts, 2002.
[48] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, N.J., 1976.
[49] P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66:191–243, 1994.
[50] D. A. Sprott. *Statistical Inference in Science*. Springer-Verlag, Berlin, 2000.
[51] Z. Tong, P. Xu, and T. Denœux. An evidential classifier based on Dempster-Shafer theory and deep learning. *Neurocomputing*, 450:275–293, 2021.
[52] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.
[53] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9(1):60–62, 1938.
[54] P. Xu, F. Davoine, H. Zha, and T. Denœux. Evidential calibration of binary SVM classifiers. *International Journal of Approximate Reasoning*, 72:55–70, 2016.
[55] L. A. Zadeh. Fuzzy sets. *Inform. Control*, 8:338–353, 1965.
[56] L. A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning –I. *Information Sciences*, 8:199–249, 1975.
[57] L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.
[58] L. A. Zadeh. Fuzzy sets and information granularity. In M. M. Gupta, R. K. Ragade, and R. R. Yager, editors, *Advances in Fuzzy Sets Theory and Applications*, pages 3–18. North-Holland, Amsterdam, 1979.

Appendix A. Proof of Proposition 1

From Zadeh's extension principle,

$$\widetilde{Z}(z) = \sup\{\widetilde{\beta}(\beta) : \boldsymbol{U}\beta = z\}.$$

Consequently, we need to solve the following constrained minimization problem,

$$\min_{\beta}(\beta - m)^T \boldsymbol{H}(\beta - m)$$

subject to $\boldsymbol{U}\beta = z$. The Lagrange function is

$$\mathcal{L}(\beta, \lambda) = (\beta - m)^T \boldsymbol{H}(\beta - m) - (\boldsymbol{U}\beta - z)^T \lambda,$$

where $\lambda \in \mathbb{R}^q$ is a q-dimensional vector of Lagrange multipliers. Its gradient w.r.t. β is

$$\frac{\partial \mathcal{L}}{\partial \beta} = 2\boldsymbol{H}(\beta - m) - \boldsymbol{U}^T\lambda.$$

Setting $\frac{\partial \mathcal{L}}{\partial \beta} = 0$ gives us

$$\beta^* = m + \frac{1}{2}\boldsymbol{H}^{-1}\boldsymbol{U}^T\lambda. \tag{A.1}$$

From $\boldsymbol{U}\beta^* = z$, we get

$$\boldsymbol{U}m + \frac{1}{2}\boldsymbol{U}\boldsymbol{H}^{-1}\boldsymbol{U}^T\lambda = z,$$

or

$$\boldsymbol{U}\boldsymbol{H}^{-1}\boldsymbol{U}^T\lambda = 2(z - \boldsymbol{U}m).$$

Now, as $\boldsymbol{H}$ is positive definite, $\text{rank}(\boldsymbol{U}\boldsymbol{H}^{-1}\boldsymbol{U}^T) = \text{rank}(\boldsymbol{U}) = q$ (See [43, p. 51]), hence $\boldsymbol{U}\boldsymbol{H}^{-1}\boldsymbol{U}^T$ is regular. Consequently, we have

$$\lambda = 2(\boldsymbol{U}\boldsymbol{H}^{-1}\boldsymbol{U}^T)^{-1}(z - \boldsymbol{U}m). \tag{A.2}$$

From (A.1) and (A.2), we get

$$\beta^* = m + \boldsymbol{H}^{-1}\boldsymbol{U}^T(\boldsymbol{U}\boldsymbol{H}^{-1}\boldsymbol{U}^T)^{-1}(z - \boldsymbol{U}m).$$

Finally, we obtain $\widetilde{Z}(z)$ as

$$\begin{aligned}
\widetilde{Z}(z) &= \exp\left(-\frac{1}{2}(\beta^* - m)^T\boldsymbol{H}(\beta^* - m)\right) \\
&= \exp\left(-\frac{1}{2}(z - \boldsymbol{U}m)^T(\boldsymbol{U}\boldsymbol{H}^{-1}\boldsymbol{U}^T)^{-1}\boldsymbol{U}\boldsymbol{H}^{-1}\boldsymbol{H}\boldsymbol{H}^{-1}\boldsymbol{U}^T(\boldsymbol{U}\boldsymbol{H}^{-1}\boldsymbol{U}^T)^{-1}(z - \boldsymbol{U}m)\right) \\
&= \exp\left(-\frac{1}{2}(z - \boldsymbol{U}m)^T(\boldsymbol{U}\boldsymbol{H}^{-1}\boldsymbol{U}^T)^{-1}(z - \boldsymbol{U}m)\right).
\end{aligned}$$

Appendix B. Related work

In this section, we briefly compare evidential likelihood-based inference recalled in Section 2.3 to other methods for quantifying uncertainty in statistical inference using belief functions. The first such method was introduced by Dempster [8][9]. Dempster's approach starts with a φ-equation such as (10), and a probability distribution P_U of $U \in \mathbb{U}$. Having observed $\boldsymbol{Y} = \boldsymbol{y}$, Dempster assumes that the analyst "continues to believe" that the uncertainty on U is quantified by P_U, and considers the random set $U \to \Gamma_{\boldsymbol{y}}(U) = \{\theta \in \Theta : \boldsymbol{y} = \varphi(\theta, U)\}$, which defines the following belief function on $\boldsymbol{\theta}$:

$$Bel_{\boldsymbol{y}}(A) = P_U\left(\emptyset \neq \Gamma_{\boldsymbol{y}}(U) \subseteq A\right) \tag{B.1}$$

for any measurable $A \subseteq \Theta$. This belief function satisfies Requirements R_1 and R_2 stated in Section 2.3 but, as shown in [13], it is more committed (in some sense) than the likelihood-based belief function (8). The calculation of degrees of belief using (B.1) also poses severe

technically difficulties except for very simple statistical models, a limitation that has hindered the application of Dempster's approach until now.

In [41], Martin and Liu argue against the "continue to believe" assumption and proposed to predict U, after observing $\boldsymbol{y}$, by a predictive random set $U \rightarrow \mathcal{S}(U)$, where $\mathcal{S}$ is a strongly measurable mapping from $\mathbb{U}$ to $2^{\mathbb{U}}$. They consider the multi-valued mapping $U \rightarrow \Gamma_{\boldsymbol{y}}^{\mathcal{S}} = \bigcup_{v \in \mathcal{S}(U)} \Gamma_{\boldsymbol{y}}(v)$, and they design the predictive random set $\mathcal{S}$ in such a way that the belief function $A \rightarrow Bel_{\boldsymbol{y}}^{\mathcal{S}}(A) = P_U\left(\emptyset \neq \Gamma_{\boldsymbol{y}}^{\mathcal{S}}(U) \subseteq A\right)$ is "valid", in the sense that, for all measurable $A \subseteq \Theta$ and all $\alpha \in (0,1)$,

$$\sup_{\theta \notin A} P_{\boldsymbol{Y}|\theta}(Bel_{\boldsymbol{Y}}^{\mathcal{S}}(A) \geq \alpha) \leq \alpha. \tag{B.2}$$

Martin and Liu call this method for constructing a belief function $Bel_Y^{\mathcal{S}}$ an *Inferential Model* (IM). In recent work [40], Martin proposes another IM construction that consists in transforming the relative likelihood (7) into another possibility distribution $\pi'_{\boldsymbol{\theta}|\boldsymbol{y}}$ defined as

$$\pi'_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta) = P_{Y|\theta}\left(\pi_{\boldsymbol{\theta}|\boldsymbol{Y}}(\theta) \leq \pi_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta)\right).$$

As shown in [40], $\pi'_{\boldsymbol{\theta}|\boldsymbol{y}}$ verifies the following "strong validity" property,

$$\forall \alpha \in [0,1], \quad \sup_{\theta \in \Theta} P_{\boldsymbol{Y}|\theta}\left(\pi'_{\boldsymbol{\theta}|\boldsymbol{Y}}(\theta) \leq \alpha\right) \leq \alpha, \tag{B.3}$$

which implies that the corresponding necessity measure (a belief function) verifies the validity property (B.2). An in-depth discussion of this method can be found in [3].

Possibility distribution $\pi'_{\boldsymbol{\theta}|\boldsymbol{y}}$ does not satisfy requirements R_1 and R_2: in particular, a "valid" possibility distribution computed from two independent samples cannot be obtained by combining the valid possibility distributions from each of the two samples using a formal rule such as the product-intersection operator. Conversely, the pure likelihood-based possibility distribution $\pi_{\boldsymbol{\theta}|\boldsymbol{y}}$ does not verify property (B.3) in general, even asymptotically. Indeed, from Wilks' theorem [53], we know that, for a independent sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ of size n, under some regularity conditions, $-2\ln \pi_{\boldsymbol{\theta}|\boldsymbol{Y}}(\theta_0)$ converges in distribution to a chi square distribution with p degrees of freedom, where θ_0 is the true value of the parameter and p is the dimension of $\boldsymbol{\theta}$; consequently,

$$\lim_{n \rightarrow \infty} P_{\boldsymbol{Y}|\theta}\left(\pi_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta) \leq \alpha\right) = 1 - F_{\chi_p^2}(-2\ln \alpha).$$

For any $\alpha \in (0,1)$, $1 - F_{\chi_p^2}(-2\ln\alpha) \leq \alpha$ if $p \leq 2$, but $1 - F_{\chi_p^2}(-2\ln\alpha) > \alpha$ if $p > 2$. However, as shown in [37], the belief function $Bel_{\boldsymbol{\theta}|\boldsymbol{y}}$ computed from $\pi_{\boldsymbol{\theta}|\boldsymbol{y}}$ by (8) is consistent in the sense that, under mild conditions, for any neighborhood $\mathcal{N}$ of θ_0, $Bel_{\boldsymbol{\theta}|\boldsymbol{y}}(\mathcal{N}) \rightarrow 1$ and $Bel_{\boldsymbol{\theta}|\boldsymbol{y}}(\mathcal{N}^c) \rightarrow 0$ almost surely under the law determined by θ_0.

Possibility distributions $\pi_{\boldsymbol{\theta}|\boldsymbol{y}}$ and $\pi'_{\boldsymbol{\theta}|\boldsymbol{y}}$ thus meet different requirements: $\pi'_{\boldsymbol{\theta}|\boldsymbol{y}}$ is frequency-calibrated in the sense of (B.3), which makes it a powerful tool for computing exact confidence regions and designing efficient testing procedures. In contrast, $\pi_{\boldsymbol{\theta}|\boldsymbol{y}}$ verifies requirements R_1 and R_2 and, as such, it can be argued to be more suitable for uncertain reasoning; in

particular, it can be directly combined with possibility measures representing other independent statistical evidence or prior information (in that sense, it is a complete representation of statistical evidence). The choice between these two solutions depends on how much importance one attaches to frequentist properties versus a simple calculus for uncertain reasoning by evidence aggregation as originally envisioned by Dempster [8] and Shafer [48]. This choice problem may ultimately be somewhat analogous to the frequentist-Bayesian dilemma, i.e., long-run frequentist properties versus a simple and well-founded mechanism for combining data with probabilistic prior knowledge.