Imprecise probability models for inference in exponential families

Erik Quaeghebeur
SYSTeMS research group,
EESA department, Ghent University
Technologiepark-Zwijnaarde 914,
9052 Zwijnaarde, Belgium
Erik.Quaeghebeur@UGent.be

Gert de Cooman
SYSTeMS research group,
EESA department, Ghent University
Technologiepark-Zwijnaarde 914,
9052 Zwijnaarde, Belgium
Gert.deCooman@UGent.be

Abstract

When considering sampling models described by a distribution from an exponential family, it is possible to create two types of imprecise probability models. One is based on the corresponding conjugate distribution and the other on the corresponding predictive distribution. In this paper, we show how these types of models can be constructed for any (regular, linear, canonical) exponential family, such as the centered normal distribution.

To illustrate the possible use of such models, we take a look at credal classification. We show that they are very natural and potentially promising candidates for describing the attributes of a credal classifier, also in the case of continuous attributes.

Keywords. Exponential family, Imprecise probability models, Inference, Conjugate analysis, Naive credal classifier.

1 Introduction

The imprecise Dirichlet model [11] and the imprecise Dirichlet-Multinomial model [13] were introduced as imprecise probability models for making inferences from categorical data. These models have two features of interest. They are elicited using i.i.d. samples and the parameters of the distributions they are based upon, correspond to some sort of average sample. This last feature allows for imprecision by making a particular use of pseudocounts.

The basis for these features is not only present in the case of categorical data, but also in other common sampling models, such as normal sampling. In fact, it is possible to construct similar imprecise probability models for sampling from a distribution that belongs to an exponential family. This is the main theme of this paper. So we start by introducing the exponential families of distributions in Section 2. In Section 3 we show how to construct the corresponding imprecise probability models.

The underlying ideas of the development in these two sections are the following:

- We restrict ourselves to nicely behaving sampling models, namely those described by exponential families of distributions. (Section 2.1)
- We wish to make assessments about the parameters of such a sampling model and update these assessments in the light of new information. For this, we use conjugate distributions, so that the prior and posterior (obtained after updating the prior) belong to the same class. The general expression for the conjugate can be given. (Section 2.2)
- We also wish to make predictive statements about future samples. A predictive distribution can easily be given using the expression of the exponential family and its conjugate. (Section 2.3)
- Lack of specific prior information leads to the use of imprecise probability models.
- Both the conjugate and the predictive distribution are parameterized by: (Section 3.1)
  (i) a parameter s that can be made to vary in a set $\mathcal{S}$ (which initially, before updating, is chosen to reflect the prior information), producing a coherent lower prevision by using the lower envelope theorem;
  (ii) a parameter n acting like sample counts, whose initially chosen value (pseudocounts) determines how fast the imprecision is reduced by updating.

The imprecise Dirichlet model can be used for constructing the naive credal classifier [14], which does a classification on the basis of categorical (discrete) attributes. Using the models we introduce in this paper, we show in Section 4 that the naive credal classifier can be extended to allow for continuous attributes.

2 Exponential families

Let us give a summary of the relevant theory about exponential families. As this is only a partial overview, we refer to the literature [7, 5, 1] for more detailed information. The theoretical exposition is interspersed with a simple but representative example, illustrating the theoretical concepts we introduce.

2.1 An exponential family

We look at sampling models where i.i.d. samples of a random variable (or vector) X are taken from a sample space $\mathcal{X}$ that is distributed according to an *exponential family*.[1] Such a distribution can be defined by giving its probability (density or mass) function

$$\mathsf{Ef}(x \mid \psi) = \mathsf{a}(x) \exp(\langle \psi, \tau(x) \rangle - \mathsf{b}(\psi)),\ x \in \mathcal{X}. \qquad (1)$$

In this expression, $\tau : \mathcal{X} \to \mathcal{T}$ is a so-called *sufficient statistic* of X (more about this in Section 2.4.1) and $\psi \in \Psi$ is a so-called *canonical parameter*. Both $\mathcal{T}$ and Ψ are (subsets of) finite-dimensional real vector spaces and $\langle \cdot, \cdot \rangle$ is a scalar product between elements of these subsets. Particular to each family are the functions $\mathsf{a} : \mathcal{X} \to \mathbb{R}^+$ and $\mathsf{b} : \Psi \to \mathbb{R}$.[2]

As an example, we look at the centered normal distribution. This is a relatively simple case, but the calculations are still representative of what is necessary for other families. To obtain the form of Equation (1) *we rewrite its classical probability density function:*

$$\begin{aligned}
\mathsf{N}(x \mid 0, \sigma) &= \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{x^2}{2\sigma^2}) \\
&\qquad\qquad (\textit{with } x \in \mathbb{R} = \mathcal{X},\ \sigma \in \mathbb{R}^+) \\
&= \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2\sigma^2} x^2 - \ln(\sigma)) \\
&= \frac{1}{\sqrt{2\pi}} \exp(\psi \tau(x) + \frac{1}{2} \ln(-2\psi)) \\
&\qquad (\textit{with } \tau(x) = x^2 \in \mathbb{R}_0^+ = \mathcal{T},\ \psi = -\frac{1}{2\sigma^2} \in \mathbb{R}^- = \Psi)
\end{aligned}$$

We can see that for this example, the scalar product is an algebraic product, $\mathsf{a} = 1/\sqrt{2\pi}$, *and* $\mathsf{b}(\psi) = -\ln(-2\psi)/2$.

A nice property of these distributions is that $P(\tau \mid \psi) = \nabla \mathsf{b}$. Here, we introduced our notation for the linear prevision (expectation) associated with the distribution considered. It is defined as follows:

$$P(f \mid \psi) = \int_{\mathcal{X}} \mathsf{Ef}(\cdot \mid \psi) f,$$

where $\int_{\mathcal{X}}$ stands for integration or summation over the space $\mathcal{X}$, and f is an element of $\mathcal{L}(\mathcal{X})$, the set of measurable gambles (bounded functions) on $\mathcal{X}$. (Note: we use similar terminology and notation further on.)

In our example, $P(X^2 \mid \psi) = \nabla \mathsf{b} = -1/2\psi$, *which is (evidently) equal to the variance* σ^2 *of the centered normal distribution.*

2.2 The conjugate distribution

When reinterpreting the probability function in Equation (1) as a likelihood function

$$\mathsf{L}_x : \Psi \to \mathbb{R}^+ : \psi \to \mathsf{Ef}(x \mid \psi),$$

we can define the *corresponding conjugate distribution* [5, 1] by giving its probability density function

$$\mathsf{CEf}(\psi \mid n, y) = \mathsf{c}(n, y) \exp(n\,[\langle \psi, y \rangle - \mathsf{b}(\psi)]),\ \psi \in \Psi. \quad (2)$$

There are two parameters, n and y. The first, $n \in \mathbb{R}^+$, can be interpreted as a number of counts (possibly including some so-called *pseudocounts*). The other, $y \in \mathcal{Y}$, corresponds to an average sufficient statistic, so it is natural that $\mathcal{Y}$ is the convex hull $\mathrm{co}(\mathcal{T})$ of $\mathcal{T}$ without—for technical reasons—the border. The function c represents a normalization factor.

A *prior* distribution with density $\mathsf{CEf}(\cdot \mid n, y)$ can be *updated* after observing a sample x. This gives a *posterior* distribution with density $p(\cdot \mid n, y, x) \propto \mathsf{CEf}(\cdot \mid n, y)\mathsf{L}_x$. This posterior's density is equal to $\mathsf{CEf}(\cdot \mid n+1, \frac{ny+\tau(x)}{n+1})$ and thus a member of the same class as the prior. This property is called *conjugacy*.

We now have enough information to find the conjugate distribution for our example. From $\mathcal{T} = \mathbb{R}_0^+$ *we derive that* $\mathcal{Y} = \mathbb{R}^+$. *To determine the normalization function* c, *we transform* Ψ *such that* ψ *is mapped to the so-called precision* $\lambda = \frac{1}{\sigma^2} = -2\psi$:

$$\begin{aligned}
\mathsf{CEf}(\psi \mid n, y) \mathrm{d}\psi &= \mathsf{c}(n, y) \exp(n \left[-\frac{\lambda}{2} y + \frac{1}{2} \ln(\lambda) \right]) \left\| \frac{\mathrm{d}\psi}{\mathrm{d}\lambda} \right\| \mathrm{d}\lambda \\
&= \frac{1}{2} \mathsf{c}(n, y) \lambda^{\frac{n}{2}} \exp(-\frac{ny}{2} \lambda) \mathrm{d}\lambda \\
&\propto \mathsf{Ga}(\lambda \mid \frac{n+2}{2}, \frac{ny}{2}) \mathrm{d}\lambda.
\end{aligned}$$

This allows us to use the normalization factor $\beta^\alpha / \Gamma(\alpha)$ *of the probability density function* $\mathsf{Ga}(\cdot \mid \alpha, \beta)$ *of the gamma distribution to find*

$$\mathsf{c}(n, y) = 2 \frac{\left[\frac{ny}{2} \right]^{\frac{n+2}{2}}}{\Gamma(\frac{n+2}{2})},$$

where Γ *is the gamma function.*

[1] To be more precise and to follow the nomenclature in the literature [7, 1], we should say: *regular, linear, canonical exponential family*.

[2] Notation: $\mathbb{R}^+$ is the set of strictly positive reals. Further on, we use $\mathbb{R}_0^+$, the set of nonnegative reals, and $\mathbb{N}_0$, the set of nonnegative integers.

Also illustrated in the above example is the following general idea. By applying a transformation to the parameter space Ψ that maps ψ to an element of the classical parameter space of the exponential family considered, the conjugate can usually be written in terms of well-known density functions. Besides helping interpretation, this also leads to an easy way of determining the normalization function c.

A nice property of the conjugate prevision $P_{\mathsf{C}}(\cdot\,|\,n,y)$ on $\mathcal{L}(\Psi)$ associated with a conjugate distribution, is that $P_{\mathsf{C}}(\nabla\mathsf{b}\,|\,n,y) = y$. This implies that $P_{\mathsf{C}}(P(\tau\,|\,\Psi)\,|\,n,y) = y$—where $P(\cdot\,|\,\Psi)$ is the function that maps ψ to $P(\cdot\,|\,\psi)$—allowing us to give an interpretation to y.

For our example, the prevision of the variance—which is clearly of interest for inference problems—can now be easily determined: $P_{\mathsf{C}}(P(\tau\,|\,\Psi)\,|\,n,y) = P_{\mathsf{C}}(\sigma^2\,|\,n,y) = y$. This tells us that y can be interpreted as a variance.

2.3 The predictive distribution

Using the conjugate distribution, we can also derive the *corresponding predictive distribution* [1]. Its probability function is given by

$$\mathsf{PEf}(x\,|\,n,y) = \int_{\Psi} \mathsf{CEf}(\cdot\,|\,n,y)\mathsf{L}_x = \frac{\mathsf{c}(n,y)\mathsf{a}(x)}{\mathsf{c}(n+1,\frac{ny+\tau(x)}{n+1})},\ x \in \mathcal{X}. \tag{3}$$

The predictive prevision associated with the predictive distribution is $P_{\mathsf{P}}(\cdot\,|\,n,y)$ on $\mathcal{L}(\mathcal{X})$.

Combining the results of the previous fragments of our example, we can write down the probability density function of the predictive distribution,

$$\mathsf{PEf}(x\,|\,n,y) = \frac{1}{\sqrt{\pi}}\frac{\Gamma(\frac{n+3}{2})}{\Gamma(\frac{n+2}{2})}\frac{[ny]^{\frac{n+2}{2}}}{[ny+x^2]^{\frac{n+3}{2}}}.$$

2.4 Remarks

2.4.1 Multiple samples

The joint distribution for m i.i.d. samples x_j is also an exponential family distribution with the same conjugate. One just applies the following changes to Equation (1):

$$\tau(x) \to \tau(x_1,\ldots,x_m) = \sum_j \tau(x_j),$$
$$\mathsf{a}(x) \to \mathsf{a}(x_1,\ldots,x_m) = \prod_j \mathsf{a}(x_j),$$
$$\mathsf{b} \to m\mathsf{b}$$

Additionally, one might have to multiply a by a factor (such as $m!$) due to limited knowledge about the ordering of the samples, but for simplicity's sake, we disregard this here.

The dimension of the sufficient statistic (i.e., a statistic containing all the information in the sample that is relevant for inference) remains the same, independent of the number of samples. Exponential families of distributions are the only families for which such finite sufficient statistics exist [1].

The corresponding likelihood function $\mathsf{L}_{x_1,\ldots,x_m}$ can then also be used for updating and for calculating a predictive distribution. After updating $\mathsf{CEf}(\cdot\,|\,n,y)$, we obtain $\mathsf{CEf}(\cdot\,|\,n+m,\frac{ny+\sum_j \tau(x_j)}{n+m})$. The probability function of the predictive distribution becomes

$$\mathsf{PEf}(x_1,\ldots,x_m\,|\,n,y) = \frac{\mathsf{c}(n,y)\prod_j \mathsf{a}(x_j)}{\mathsf{c}(n+m,\frac{ny+\sum_j \tau(x_j)}{n+m})}.$$

2.4.2 Reference table

The characteristics of exponential families as we describe them here are not commonly found in the literature. Therefore, we have included Table 1 for easy reference. For some common sampling models that are described by an exponential family, it contains information similar to that derived for the centered normal in our example.

3 Imprecise probability models

Some ideas for using imprecise probability models involving exponential families for inference can be found in the literature. One idea takes a prior conjugate distribution with fixed ny and uses the neighborhood around this prior created by varying ny [2] (robust Bayesian literature). Another idea uses lower and upper density functions [4] (imprecise probabilities literature).

The approach we present in this paper differs from the ones cited above, because it isn't based on lower and upper density functions and because it doesn't start from one fixed prior distribution, but uses a convex set of distributions. Our approach is inspired by the approach to inference from categorical data taken in the imprecise Dirichlet model or IDM [11] and the imprecise Dirichlet-Multinomial model or IDMM [13].

We should also mention the *bounded derivative model* [12]. This model is defined by the set of all strictly positive, continuous, smooth probability density functions that have a bounded logarithmic derivative. It is of interest because it produces tractable inferences for $P(\tau\,|\,\psi)$ when the sampling model is described by a one-parameter exponential family. This is also the case for our model, even outside of the one-parameter case. We will comment on this further on in Section 3.2, where we introduce this result.

[3] Notation in Table 1: $\mathbb{R}^{d\times d}_{\text{sy,pd}}$ are the symmetrical positive definite matrices and

$$\Gamma_d(z) = \pi^{\frac{d[d-1]}{4}}\prod_{i=1}^{d}\Gamma(\tfrac{2z+1-i}{2})$$

is the generalized gamma function.

Exponential family *probability function* *(classical parameters used)*	$\mathcal{X}$	ψ	$\tau(x)$	$\mathcal{Y}$
Normal $\mathsf{N}(x \mid \mu, \sigma), \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$ *(take $\lambda = \frac{1}{\sigma^2}$)*	$\mathbb{R}$	$\begin{pmatrix} \lambda\mu \\ -\frac{1}{2}\lambda \end{pmatrix}$	$\begin{pmatrix} x \\ x^2 \end{pmatrix}$	$\left\{y \in \mathbb{R} \times \mathbb{R}^+ : y_2 - {y_1}^2 > 0\right\}$
Centered normal $\mathsf{N}(x \mid 0, \sigma), \sigma \in \mathbb{R}^+$ *(take $\lambda = \frac{1}{\sigma^2}$)*	$\mathbb{R}$	$-\frac{1}{2}\lambda$	x^2	$\mathbb{R}^+$
Scaled normal $\mathsf{N}(x \mid \mu, 1), \mu \in \mathbb{R}$	$\mathbb{R}$	μ	x	$\mathbb{R}$
Multivariate normal[3] $\mathsf{N}(x \mid \mu, \Sigma), \mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d\times d}_{\text{sy,pd}}$ *(take $\Lambda = \Sigma^{-2}$)*	$\mathbb{R}^d$	$\begin{pmatrix} \Lambda\mu \\ -\frac{1}{2}\Lambda \end{pmatrix}$	$\begin{pmatrix} x \\ xx^T \end{pmatrix}$	$\left\{y \in \mathbb{R}^d \times \mathbb{R}^{d\times d}_{\text{sy,pd}} : y_2 - y_1{y_1}^T \in \mathbb{R}^{d\times d}_{\text{sy,pd}}\right\}$
Bernoulli $\mathsf{Br}(x \mid \theta), \theta \in (0, 1)$	$\{0, 1\}$	$\ln(\frac{\theta}{1-\theta})$	x	$(0, 1)$
Multivariate Bernoulli $\mathsf{Br}(x \mid \theta), \theta \in (0, 1)^d : \|\theta\| < 1$ *(take $\theta_0 = 1 - \sum_i \theta_i$)*	$\left\{x \in \{0, 1\}^d : \|x\| \leq 1\right\}$	$\left(\ln(\frac{\theta_i}{\theta_0})\right)_{i=1}^d$	x	$\left\{y \in (0, 1)^d : \|y\| < 1\right\}$ *(take $y_0 = 1 - \sum_i y_i$)*
Exponential $\mathsf{Ex}(x \mid \beta), \beta \in \mathbb{R}^+$	$\mathbb{R}^+_0$	$-\beta$	x	$\mathbb{R}^+$
Poisson $\mathsf{Pn}(x \mid \lambda), \lambda \in \mathbb{R}^+$	$\mathbb{N}_0$	$\ln(\lambda)$	x	$\mathbb{R}^+$

Exponential family *probability function* *(classical parameters used)*	a	b	c	∇b	Conjugate *probability density function* *(classical parameters as argument)*
Normal $\mathsf{N}(x \mid \mu, \sigma), \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$ *(take $\lambda = \frac{1}{\sigma^2}$, $m_2 = \sigma^2 + \mu^2$)*	$\frac{1}{\sqrt{2\pi}}$	$\frac{\lambda\mu^2 - \ln(\lambda)}{2}$	$\frac{2\sqrt{n}}{\sqrt{2\pi}} \frac{\left[\frac{n\left[y_2 - {y_1}^2\right]}{2}\right]^{\frac{n+3}{2}}}{\Gamma(\frac{n+3}{2})}$	$\begin{pmatrix} \mu \\ m_2 \end{pmatrix}$	Normal-gamma $\mathsf{N}(\mu \mid y_1, n\lambda)\mathsf{Ga}(\lambda \mid \frac{n+3}{2}, \frac{n\left[y_2 - {y_1}^2\right]}{2})$
Centered normal $\mathsf{N}(x \mid 0, \sigma), \sigma \in \mathbb{R}^+$ *(take $\lambda = \frac{1}{\sigma^2}$)*	$\frac{1}{\sqrt{2\pi}}$	$-\frac{\ln(\lambda)}{2}$	$\frac{2\left[\frac{ny}{2}\right]^{\frac{n+2}{2}}}{\Gamma(\frac{n+2}{2})}$	σ^2	Gamma $\mathsf{Ga}(\lambda \mid \frac{n+2}{2}, \frac{ny}{2})$
Scaled normal $\mathsf{N}(x \mid \mu, 1), \mu \in \mathbb{R}$	$\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$	$\frac{\mu^2}{2}$	$\frac{\sqrt{n}e^{\frac{ny^2}{2}}}{\sqrt{2\pi}}$	μ	Normal $\mathsf{N}(\mu \mid y, n)$
Multivariate normal[3] $\mathsf{N}(x \mid \mu, \Sigma), \mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d\times d}_{\text{sy,pd}}$ *(take $\Lambda = \Sigma^{-2}$, $M_2 = \Sigma^2 + \mu\mu^T$)*	$\frac{1}{\sqrt{2\pi}^d}$	$\frac{\mu^T\Lambda\mu - \ln(\lvert\Lambda\rvert)}{2}$	$\frac{2\sqrt{n}}{\sqrt{2\pi}^d} \frac{\left[\frac{n\left\lvert y_2 - y_1{y_1}^T\right\rvert}{2}\right]^{\frac{n+d+2}{2}}}{\Gamma_d(\frac{n+d+2}{2})}$	$\begin{pmatrix} \mu \\ M_2 \end{pmatrix}$	Normal-Wishart $\mathsf{N}(\mu \mid y_1, n\Lambda)\mathsf{Wi}(\Lambda \mid \frac{n+d+2}{2}, \frac{n\left[y_2 - y_1{y_1}^T\right]}{2})$
Bernoulli $\mathsf{Br}(x \mid \theta), \theta \in (0, 1)$	1	$\ln(1 - \theta)$	$\frac{\Gamma(n)}{\Gamma(n\left[1-y\right])\Gamma(ny)}$	θ	Beta $\mathsf{Be}(\theta \mid ny, n\left[1-y\right])$
Multivariate Bernoulli $\mathsf{Br}(x \mid \theta), \theta \in (0, 1)^d : \|\theta\| < 1$ *(take $\theta_0 = 1 - \sum_i \theta_i$)*	1	$\ln(\theta_0)$	$\frac{\Gamma(n)}{\Gamma(ny_0)\prod_i \Gamma(ny_i)}$	θ	Dirichlet $\mathsf{Di}(\theta \mid ny, ny_0)$
Exponential $\mathsf{Ex}(x \mid \beta), \beta \in \mathbb{R}^+$	1	$-\ln(\beta)$	$\frac{[ny]^{n+1}}{\Gamma(n+1)}$	$\frac{1}{\beta}$	Gamma $\mathsf{Ga}(\beta \mid n+1, ny)$
Poisson $\mathsf{Pn}(x \mid \lambda), \lambda \in \mathbb{R}^+$	$\frac{1}{x!}$	λ	$\frac{n^{ny}}{\Gamma(ny)}$	λ	Gamma $\mathsf{Ga}(\lambda \mid ny, n)$

Table 1: Characteristics of some commonly used exponential families

In this section, we define and investigate our imprecise probability models from a theoretical perspective. To make this discussion more tangible and clear, we again give an example and make the link to the already established IDM and IDMM models [11, 13].

3.1 Definitions

3.1.1 Notation

Up until now, we haven't made any special distinction between priors and posteriors. A prior could have been the posterior of another prior. For what follows it is necessary to introduce an initial prior, which is elicited on the basis of assumptions about the sampling model under study, but not on any observed samples.

We use an upper index $k \in \mathbb{N}_0$ to indicate the number of samples x_j that has been used to elicit the parameters of a model. For example, a prior conjugate prevision will thus be written as $P_{\mathsf{C}}(\cdot \,|\, n^0, y^0)$ and a predictive prevision based on k samples will be denoted by $P_{\mathsf{P}}(\cdot \,|\, n^k, y^k)$.[4]

Remembering Section 2.4.1, it is easy to see that

$$n^k = n^0 + k, \qquad y^k = \frac{n^0 y^0 + \tau^k}{n^0 + k}, \tag{4}$$

where we have used τ^k to abbreviate $\tau(x_1, \ldots, x_k)$.

To finish this notational digression, consider a subset $\mathcal{Y}^0$ of $\mathcal{Y}$. We define

$$\mathcal{Y}^k = \left\{ \frac{n^0 y + \tau^k}{n^0 + k} \,:\, y \in \mathcal{Y}^0 \right\} \subset \mathcal{Y}. \tag{5}$$

3.1.2 Conjugate and predictive models

Both imprecise probability models we associate with an exponential family are lower previsions $\underline{P}$ that are defined as lower envelopes of linear previsions P. As such, these lower previsions are coherent [10]. We also use the conjugate upper prevision $\overline{P} = -\underline{P}(-\cdot)$.[5]

The *conjugate model* is the lower envelope—taken over a set $\mathcal{Y}^k$—of a set of conjugate previsions:

$$\underline{P}_{\mathsf{C}}(\cdot \,|\, n^k, \mathcal{Y}^k) = \inf_{y \in \mathcal{Y}^k} P_{\mathsf{C}}(\cdot \,|\, n^k, y).$$

This lower prevision is defined on $\mathcal{L}(\Psi)$. Although this is possible, we will not look at the case where the lower envelope is also taken over a set of counts.

The *predictive model* is the lower envelope—again taken over a set $\mathcal{Y}^k$—of a set of predictive previsions:

$$\underline{P}_{\mathsf{P}}(\cdot \,|\, n^k, \mathcal{Y}^k) = \inf_{y \in \mathcal{Y}^k} P_{\mathsf{P}}(\cdot \,|\, n^k, y).$$

This lower prevision is defined on $\mathcal{L}(\mathcal{X})$. This model can be seen as a restriction of the first using likelihood functions, i.e., $\underline{P}_{\mathsf{P}}(f \,|\, n^k, \mathcal{Y}^k) = \underline{P}_{\mathsf{C}}(\int_{\mathcal{X}} f(x) \mathsf{L}_x \mathrm{d}x \,|\, n^k, \mathcal{Y}^k)$, where $f \in \mathcal{L}(\mathcal{X})$.

The *credal sets* corresponding to the models given above consist of the closure of convex mixtures of the distributions corresponding to the respective probability functions $\mathsf{CEf}(\cdot \,|\, n^k, y)$ and $\mathsf{PEf}(\cdot \,|\, n^k, y)$, where $y \in \mathcal{Y}^k$.

3.1.3 The set $\mathcal{Y}^k$

Now let us turn our attention to the set $\mathcal{Y}^k \subset \mathcal{Y}$. In Section 3.1.1, $\mathcal{Y}^k$ is defined as a convex 'mixture' of $\mathcal{Y}^0 \subset \mathcal{Y}$ and $\tau^k/k \in \mathrm{co}(\mathcal{T})$ with respective coefficients $n^0/[n^0 + k]$ and $k/[n^0 + k]$. This tells us that $\mathcal{Y}^k$ is a translated (over $\tau^k/[n^0 + k]$) and scaled (factor $n^0/[n^0 + k]$) version of $\mathcal{Y}^0$.

The imprecision of the inferences of a conjugate or predictive model is (not necessary linearly) proportional to the volume of the convex hull of $\mathcal{Y}^k$ (relative to the volume of $\mathcal{Y}$). This indicates that the larger the number of pseudocounts $n^0 \in \mathbb{R}^+$ is, the slower the scaling factor increases, which results in a more conservative learning model. The choice of the number of pseudocounts depends on the application and is as such partly arbitrary.

The set $\mathcal{Y}^0$ should be chosen such that it reflects the initial assumptions. It will often be required that inferences from the initial prior are very conservative (expressing some form of 'near ignorance' [10]) and choosing $\mathcal{Y}^0 = \mathcal{Y}$ would seem ideal. However, to make sure that the assessments produced by the models do not remain vacuous as more observations are made, $\mathcal{Y}^0$ should be bounded. A result of this is that the imprecision decreases as more observations are made, so no dilation (see, e.g., [8]) occurs with these models. The choice of bound is again application-dependent and as such partly arbitrary. Note that it should not be hard to specify reasonable bounds, considering we have already assumed it was possible to restrict the sampling model to a specific exponential family.

As an example, let us look at the case of normal sampling. From Table 1, we see that we can choose $\mathcal{Y}_0$ by taking a bound ${\alpha_1}^2$ for ${y_1}^2$ and a bound $\alpha_2 + {y_1}^2$ for y_2. The rationale for this choice will be made clear later. This example is shown in Figure 1, where we also show what happens when we update our model after observing a sample x.

Note that in the Bernoulli case (see Table 1) it isn't necessary to choose any bounds, as $\mathcal{Y}$ is already bounded (i.e., it is the so-called d-dimensional unit-simplex). In this case, the conjugate model $\underline{P}_{\mathsf{C}}(\cdot \,|\, n^k, \mathcal{Y}^k)$ is an IDM [11] and the predictive model $\underline{P}_{\mathsf{P}}(\cdot \,|\, n^k, \mathcal{Y}^k)$ is an IDMM [13].

[4] Comparing our notation with the one typically used for the IDM(M) [11, 13, 14], we get the following correspondences: $n^0 \leftrightarrow s$ and $y^0 \leftrightarrow t$.

[5] The word 'conjugate' used in this sentence expresses the given relationship between a lower and an upper prevision. It has nothing to do with the use of the word 'conjugate' in the rest of this paper, which refers to a relationship between prior, likelihood, and posterior.

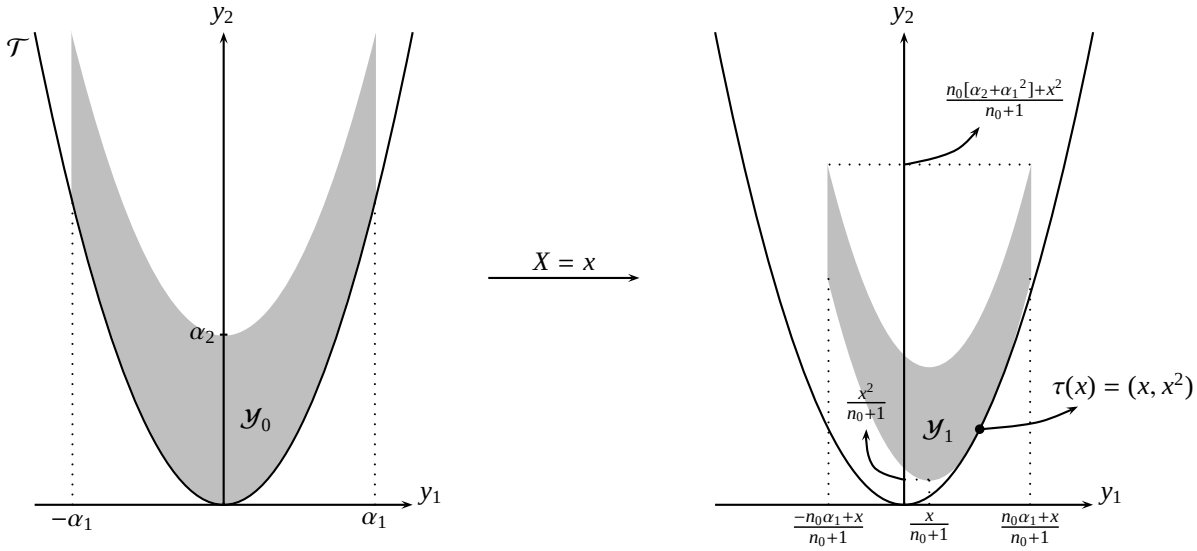

Figure 1: Case of normal sampling: choosing $\mathcal{Y}_0$ and updating to $\mathcal{Y}_1$ (both sets are colored gray).

A convex $\mathcal{Y}^0$ is used in the IDM(M), in contrast to our example. For the conjugate and predictive models, the set $\mathcal{Y}^0$ can be any continuous or discrete subset of $\mathcal{Y}$. The actual choice should be inspired by the assumptions warranted by the application at hand.

3.2 Results

We showed in Section 2.2 that for any conjugate prevision it holds that $P_{\mathsf{C}}(\nabla\mathsf{b} \mid n, y) = y$. This allows us to derive the following result for the conjugate model:

$$\begin{aligned} \underline{P}_{\mathsf{C}}(\nabla\mathsf{b} \mid n^k, \mathcal{Y}^k) &= \underline{y}^k = \frac{n^0\underline{y}^0 + ky^k}{n^0 + k}, \\ \overline{P}_{\mathsf{C}}(\nabla\mathsf{b} \mid n^k, \mathcal{Y}^k) &= \overline{y}^k = \frac{n^0\overline{y}^0 + ky^k}{n^0 + k}. \end{aligned} \tag{6}$$

Here, $\underline{y}^k$ and $\overline{y}^k$ are the pointwise infimum and supremum values of the elements of $\mathcal{Y}^k$. Because $\nabla\mathsf{b} = P(\tau \mid \psi)$ is often—though not always—a quantity of interest (see Table 1), this result shows that the calculation of inferences for these quantities is very straightforward.

We've already mentioned that for the bounded derivative model [12], a similar result holds. This is due to the fact that the credal set for the bounded derivative model includes some conjugate distributions (but, in contrast to our model, also many non-conjugate ones) and that two of these determine the upper and lower prevision of $P(\tau \mid \psi)$ (except when almost no samples have been observed).

Returning to our example of normal sampling, we know it holds that $\nabla\mathsf{b} = (\mu, m_2)$*. So after one observation, the lower and upper previsions for the mean* μ *and the noncentral second moment* m_2 *become:*

$$\underline{P}_{\mathsf{C}}((\mu, m_2) \mid n^1, \mathcal{Y}^1) = (\frac{-n^0\alpha_1 + x}{n^0 + 1}, \frac{x^2}{n^0 + 1}),$$
$$\overline{P}_{\mathsf{C}}((\mu, m_2) \mid n^1, \mathcal{Y}^1) = (\frac{n^0\alpha_1 + x}{n^0 + 1}, \frac{n^0[\alpha_2 + \alpha_1{}^2] + x^2}{n^0 + 1}).$$

These values are also indicated on Figure 1.

The fact that the values on the y_1 *axis can be interpreted as means and the values on the* y_2 *as a noncentral second moments is what led us to our choice of bounds. Choosing* $\alpha_2 + {y_1}^2$ *as a bound for* y_2 *is a seemingly reasonable ad-hoc way of bounding the variance, because* $m_2 = \sigma^2 + \mu^2$*. The need to take a bounded* $\mathcal{Y}_0$ *is immediately clear: if* $\|\alpha\| \to +\infty$*, up to three out of four of the above inferences would remain unchanged, no matter how many samples we observe. This is clearly unwanted behavior.*

If we take the difference between upper and lower prevision as a measure for the imprecision, we see that in this example, after m observations, we get

$$\frac{n^0}{n^0 + m}(2\alpha_1, \alpha_2 + {\alpha_1}^2).$$

This illustrates the remark made earlier about the learning conservatism increasing with n^0*. We see it takes* $m = n^0$ *observations to decrease the imprecision to half its initial value.*

A similarly general result as in Equation (6) for the predictive model seems unlikely given the large variation in functional form of the probability function (3) on which it is based (see the functions a and c in Table 1).

However, it is useful to cite a nice property for the predictive model when the sampling distribution is a multivariate Bernoulli (see Table 1), which is the single-sample version of the multinomial distribution. This predictive model

is an IDMM.[6] For the linear previsions determining the lower envelope $\underline{P}_{\mathsf{P}}(\cdot \mid n^k, \mathcal{Y}^k)$, it can be shown that

$$P_{\mathsf{P}}(I_i \mid n^k, y) = y_i, \ \forall i \in \{0, \ldots, d\}, \ \forall y \in \mathcal{Y}^k, \qquad (7)$$

where I_i is the indicator function for category number i. This property is the basis for the so-called *representation invariance principle* [11], as it allows categories to be pooled.

4 Credal classification

The *naive credal classifier* or NCC [14] was constructed for classifying on the basis of one or more categorical attributes. This means that for continuous attributes (such as weight, length, etc.) a discretization must be performed. We present an approach with which it is (at least theoretically) possible to classify using the continuous attributes directly if they are distributed according to an exponential family. Note that for the naive Bayes classifier—the analogous classifier in a precise probability framework—there already exist approaches using continuous attributes [6, 3].

We first reintroduce the concept of a credal classifier, but in a different manner than in [14], in order to make our contribution fit more naturally. Again, we give a small example to illustrate the theory.

4.1 Classifying

Consider some attributes taking values in a set $\mathcal{A}$, and a set of classes C. A classifier is a function that maps attribute values $a \in \mathcal{A}$ to one or more classes $c \in C$. For example, a parent choosing a T-shirt size (the classes: small, medium, large) for a child (with attributes: size, growth rate, etc.) is a classifier.

A *credal classifier* uses a conditional imprecise probability model $\underline{P}(\cdot \mid \mathcal{A})$ defined on $\mathcal{L}(C)$ to determine the expected utility of deciding between one class and another for a given set of attribute values. The specific approach to decision making we use here is called *maximality* [10, 9]. Consider the utility functions $f_{c'}, f_{c''} \in \mathcal{L}(C)$ associated with the actions of choosing c' or choosing c''. Given attribute values a, if the lower expected utility of choosing a class c' over c'' is strictly positive, then class c' is preferred to c''. Formally:

$$\underline{P}(f_{c'} - f_{c''} \mid a) > 0 \Leftrightarrow c' \succ c''.$$

This criterion creates a strict partial order on the set of classes C. The maximal, i.e., undominated, elements of this partial order will be the output of the credal classifier.

To simplify matters, we use an indicator function I_c as the utility function for choosing a class c. This corresponds to a choice of T-shirt size being either right or wrong. It means we disregard, e.g., the fact that a T-shirt that is too large, might one day fit the growing child, but that a T-shirt that is too small, will never fit that child. Our criterion becomes

$$\underline{P}(I_{c'} - I_{c''} \mid a) > 0 \Leftrightarrow c' \succ c''. \qquad (8)$$

Of course, to use this criterion, we need the model $\underline{P}(\cdot \mid \mathcal{A})$. The construction of models that allow us to apply the above criterion is the subject of the next section.

[6] To be exact, it differs slightly, but this difference is irrelevant.

4.2 Class and attribute models

4.2.1 The general approach

First consider a *class model* that describes the knowledge about the classes. This model could—in our T-shirt example—contain the information that at least half of the children need a size medium (this is what the parent believes). For this model we use a lower prevision $\underline{P}$ on $\mathcal{L}(C)$.

Next, consider an *attribute model* that describes the knowledge about the attribute values for a given class. A parent could, e.g., believe that children that need size medium T-shirts are mostly male pre-teens. For this model, we use a conditional lower prevision $\underline{P}(\cdot \mid C)$ on $\mathcal{L}(\mathcal{A})$.

Using marginal extension [10], we combine the class model $\underline{P}$ and attribute model $\underline{P}(\cdot \mid C)$ into a *class-attribute* model $\underline{E}$ defined on $\mathcal{L}(C \times \mathcal{A})$. Explicitly, for a gamble $f \in \mathcal{L}(C \times \mathcal{A})$ we get

$$\underline{E}(f) = \underline{P}(\underline{P}(f \mid C)) = \underline{P}(\sum_{c \in C} I_c \underline{P}(f(c, \cdot) \mid c)).$$

This joint model could for instance tell us that size large T-shirt-wearing toddlers make up less than one-tenth of all the T-shirt-wearing children.

4.2.2 Specifying the models

To arrive at the probabilistic model we use in the NCC, we now specify class and attribute models. Although we know full well that other options are imaginable, we will restrict ourselves to models of the type specified in Section 3.1.2, because they form a natural generalization of the model commonly used [14].

The sample space $\mathcal{X}$ for the *class model* consists of the finite number of classes in C. Together with the fact that we suppose our samples are i.i.d., it is evident that we use a model for the multivariate Bernoulli case. As our class model must be defined on $\mathcal{L}(C)$, we have to use a predictive model $\underline{P}_{\mathsf{P}}(\cdot \mid n_C, \mathcal{Y}_C)$. (Note: to alleviate the notation we omit—wherever possible—the superscript for the number of samples used to train our model. So n_C and $\mathcal{Y}_C$ should be read as n_C^k and $\mathcal{Y}_C^k$.) The initial prior we use is based on the set $\mathcal{Y}_C^0 = \{y \in (0,1)^d : \sum_{c \in C} y_c < 1\}$. As mentioned

earlier, this model is an IDMM. Remember that the choice of initial counts n_C^0 depends on the actual application.

Incorporating our choice of class model, we can rewrite our class-attribute model for any $f \in \mathcal{L}(C \times \mathcal{A})$,

$$\begin{aligned}\underline{E}(f) &= \underline{P}_{\mathsf{P}}(\sum_{c\in C} I_c \underline{P}(f(c,\cdot)\,|\,c)\,|\,n_C, \mathcal{Y}_C)\\ &= \inf_{y\in\mathcal{Y}_C} P_{\mathsf{P}}(\sum_{c\in C} I_c \underline{P}(f(c,\cdot)\,|\,c)\,|\,n_C, y)\\ &= \inf_{y\in\mathcal{Y}_C} \sum_{c\in C} P_{\mathsf{P}}(I_c\,|\,n_C, y)\underline{P}(f(c,\cdot)\,|\,c)\\ &= \inf_{y\in\mathcal{Y}_C} \sum_{c\in C} y_c \underline{P}(f(c,\cdot)\,|\,c),\end{aligned}$$

where we used Equation (7) in the last step.

The sample space $\mathcal{X}$ for the *attribute model* is the set of attribute values $\mathcal{A}$. We assume from now on that the attribute values are distributed according to an exponential family. Given a class c, we can then use a type-1 product [10] of predictive models $\underline{P}_{\mathsf{P}}(\cdot\,|\,n_{\mathcal{A}|c}, \mathcal{Y}_{\mathcal{A}|c})$—one for every attribute—as our attribute model. Such a type-1 product can be used under the assumption that given the class, the different attributes are independent, which is why the name *naive* credal classifier is used. To simplify the notation, we will from now on suppose we only use one attribute. The generalization to multiple attributes is straightforward, although coping with the corresponding increase in computational complexity is much less so. Again, the initial parameters $n^0_{\mathcal{A}|c}$ and $\mathcal{Y}^0_{\mathcal{A}|c}$ are application-dependent and can as such be chosen relatively freely.

When taking $\underline{P}_{\mathsf{P}}(\cdot\,|\,n_{\mathcal{A}|c}, \mathcal{Y}_{\mathcal{A}|c})$ to be an IDMM, the resulting classifier corresponds to the classical definition of the NCC [14].

Taking into account the restriction of our attribute models to predictive models for exponential families, our class-attribute model can be written as ($f \in \mathcal{L}(C \times \mathcal{A})$)

$$\underline{E}(f) = \inf_{y\in\mathcal{Y}_C} \sum_{c\in C} y_c \underline{P}_{\mathsf{P}}(f(c,\cdot)\,|\,n_{\mathcal{A}|c}, \mathcal{Y}_{\mathcal{A}|c}).$$

It is useful to have a short look at how updating works in our model. This updating corresponds to the so-called *training* of our model with a set of pre-classified attribute samples, or couples of the form (c, a). (Training would—in our T-shirt example—correspond to the parent assimilating the specifics of any child with nicely fitting T-shirt that they see.) We suppose that we've already updated with a number of samples and now observe (c', a'). We update $\underline{E}$ by updating the parameters of the class and attribute models that compose it (see Equation (4)):

$$\begin{aligned} n_C &\to n_C + 1,\\ \mathcal{Y}_C &\to \left\{\left(\frac{n_C y_c + \delta_{cc'}}{n_C + 1}\right)_{c\in C} : y \in \mathcal{Y}_C\right\},\\ n_{\mathcal{A}|c'} &\to n_{\mathcal{A}|c'} + 1,\\ \mathcal{Y}_{\mathcal{A}|c'} &\to \left\{\frac{n_{\mathcal{A}|c'} y_{\mathcal{A}|c'} + a'}{n_{\mathcal{A}|c'} + 1} : y_{\mathcal{A}|c'} \in \mathcal{Y}_{\mathcal{A}|c'}\right\}.\end{aligned}$$

All the other parameters remain unchanged.

From the above it follows that $n_C y_c \to n_C y_c + \delta_{cc'}$.[7] Given we have some freedom in choosing $n_{\mathcal{A}|c}$, this property allows us to set $n_{\mathcal{A}|c} = n_C y_c$ for all $c \in C$. This is also done implicitly in the classical definition of the NCC [14].[8] Although it is possible to use values for $n_{\mathcal{A}|c}$ that do not depend on y_c (which even leads to easier calculations), the above choice allows for a very nice interpretation. We can now interpret n_C^0 as pseudocounts: a number of hypothetical observations (c, a) we use in our model. These hypothetical observations have an average sufficient statistic that can take on any value in $\mathcal{Y}_C^0 \times \mathcal{Y}_{\mathcal{A}|c}^0$. They correspond to y in Equation (5) and account for all the imprecision in our inferences. As the number of real observations (c, a) grows, the relative weight of the pseudocounts will diminish, and with it the imprecision.

So finally, the class-attribute model we are going to use can be written as ($f \in \mathcal{L}(C \times \mathcal{A})$)

$$\begin{aligned}&\underline{E}(f\,|\,n_C, \mathcal{Y}_C, \mathcal{Y}_{\mathcal{A}|C})\\ &\quad= \inf_{y\in\mathcal{Y}_C} \sum_{c\in C} y_c \underline{P}_{\mathsf{P}}(f(c,\cdot)\,|\,n_C y_c, \mathcal{Y}_{\mathcal{A}|c})\\ &\quad= \inf_{\substack{y\in\mathcal{Y}_C\\ y_{\mathcal{A}|C}\in\mathcal{Y}_{\mathcal{A}|C}}} \sum_{c\in C} y_c P_{\mathsf{P}}(f(c,\cdot)\,|\,n_C y_c, y_{\mathcal{A}|c})\\ &\quad= \inf_{\substack{y\in\mathcal{Y}_C\\ y_{\mathcal{A}|C}\in\mathcal{Y}_{\mathcal{A}|C}}} \sum_{c\in C} y_c \int_{\mathcal{A}} f(c,\cdot)\mathsf{PEf}(\cdot\,|\,n_C y_c, y_{\mathcal{A}|c}),\end{aligned}$$

where $\mathcal{Y}_{\mathcal{A}|C} = (\mathcal{Y}_{\mathcal{A}|c})_{c\in C}$.

4.3 Classifying (bis)

We now have a joint model $\underline{E}(\cdot\,|\,n_C, \mathcal{Y}_C, \mathcal{Y}_{\mathcal{A}|C})$ defined on $\mathcal{L}(C \times \mathcal{A})$, while we need the corresponding conditional model $\underline{P}(\cdot\,|\,\mathcal{A})$ on $\mathcal{L}(C)$. Using Bayes' rule for density functions [10], we can write ($g \in \mathcal{L}(C)$)

$$\underline{P}(g\,|\,\mathcal{A}) = \inf_{\substack{y\in\mathcal{Y}_C\\ y_{\mathcal{A}|C}\in\mathcal{Y}_{\mathcal{A}|C}}} \frac{\sum_{c\in C} g(c) y_c \mathsf{PEf}(a\,|\,n_C y_c, y_{\mathcal{A}|c})}{\sum_{c\in C} y_c \mathsf{PEf}(a\,|\,n_C y_c, y_{\mathcal{A}|c})},$$

if

$$\inf_{\substack{y\in\mathcal{Y}_C\\ y_{\mathcal{A}|C}\in\mathcal{Y}_{\mathcal{A}|C}}} \sum_{c\in C} y_c \mathsf{PEf}(a\,|\,n_C y_c, y_{\mathcal{A}|c}) > 0. \tag{9}$$

[7] Notation: $\delta_{\alpha\beta}$ is the Kronecker delta, which is 1 when $\alpha = \beta$ and 0 otherwise.

[8] This is compatible with a sensitivity analysis interpretation [10].

Whenever Condition (9) holds, we can rewrite Criterion (8) as follows:

$$\inf_{\substack{y\in\mathcal{Y}_C \\ y_{\mathcal{A}|C}\in\mathcal{Y}_{\mathcal{A}|C}}} [y_{c'}\mathsf{PEf}(a\,|\,n_C y_{c'}, y_{\mathcal{A}|c'}) - y_{c''}\mathsf{PEf}(a\,|\,n_C y_{c''}, y_{\mathcal{A}|c''})] > 0 \Leftrightarrow c' \succ c''.$$

This criterion can be put into its final form by realizing that the parameters $y_{\mathcal{A}|c}, c \in C$ are independent. We find

$$\inf_{y\in\mathcal{Y}_C}\left[y_{c'}\inf_{y_{\mathcal{A}|c'}\in\mathcal{Y}_{\mathcal{A}|c'}}\mathsf{PEf}(a\,|\,n_C y_{c'}, y_{\mathcal{A}|c'}) - y_{c''}\sup_{y_{\mathcal{A}|c''}\in\mathcal{Y}_{\mathcal{A}|c''}}\mathsf{PEf}(a\,|\,n_C y_{c''}, y_{\mathcal{A}|c''})\right] > 0 \Leftrightarrow c' \succ c''. \tag{10}$$

This criterion can also be shown to be equivalent to Criterion (8) if Condition (9) does not hold.

As an example, we will look at the case where the attribute values are distributed according to a centered normal distribution. Using previous results, we know that

$$\mathsf{PEf}(a\,|\,n_C y_c, y_{\mathcal{A}|c}) \propto \frac{\Gamma(\frac{n_C y_c+3}{2})}{\Gamma(\frac{n_C y_c+2}{2})}\frac{[n_C y_c y_{\mathcal{A}|c}]^{\frac{n_C y_c+2}{2}}}{[n_C y_c y_{\mathcal{A}|c}+a^2]^{\frac{n_C y_c+3}{2}}}. \tag{11}$$

To apply Criterion (10)*, we first have to calculate the infimum and supremum of this expression over $\mathcal{Y}_{\mathcal{A}|c}$, an interval in $\mathbb{R}^+$. This can be done analytically. Then a two-dimensional $(y_{c'}, y_{c''})$ constrained $(y_{c'} + y_{c''} \leq 1)$ optimization problem needs to be solved. It can be shown that for most attribute values it is possible to reduce this to a one-dimensional problem $(y_{c'} + y_{c''} = 1)$. (Note: this is always the case for discrete attributes) When the observed attribute value a is an outlier—i.e., a^2 is much larger than the lower bound of the intervals $\mathcal{Y}_{\mathcal{A}|c}, c \in \{c', c''\}$—it might be necessary to solve the more complex two-dimensional optimization problem.*

From the above example, it is clear that by not discretizing, but rather using sampling models with continuous sample spaces, the optimization problems we need to solve become more of a challenge. In the example, we have used only one attribute. In the case of multiple attributes, the Expression (11) has to be replaced by a product of possibly non-similar factors. As long as the optimization over $\mathcal{Y}_{\mathcal{A}|c}$ for each of the attributes can be done analytically, the ensuing optimization problem will be two-dimensional. The possibility that this can be reduced to one-dimensional problems always remains, but because the two terms in Criterion (10) will be more complex expressions in y_c, this becomes less likely.

4.4 Remarks

We can make some finishing remarks about the conceptual differences and similarities between using a model for continuous variables or using a model for discrete (discretized) variables, i.e., an IDM(M).

When discretizing, one can approximate any type of distribution, while the models we present are currently limited to exponential families.

What one loses during discretization however, is that the different classes may correspond to neighboring or distant parts of the sample space. (One could imagine ad-hoc ways of alleviating this problem by spreading out samples over different classes.)

The models for the attributes, given different classes, might be very different. When discretizing, this poses no problem. When using models for continuous variables, this may be taken into account by using different sampling models for different classes. One could for example take a model for centered normal variables for class c', but one for scaled normal variables for c''.

5 Conclusions

In this paper we have first looked at exponential families and the corresponding conjugate and predictive families. The manner in which these families are described allowed us to introduce two imprecise probability models for inference in exponential families. The first, the conjugate model, leads to an easy way of generating inferences about the classical parameters of the exponential family under study. The second, the predictive model, does not seem such a good candidate for obtaining general results. However, it does seem a very natural model for applications. One of these applications is the naive credal classifier, which we introduced using an approach different from the classical one, to allow for continuous attributes.

Throughout this paper, loose ends were inevitably left dangling. Some of them are irritating, some of them are promising. We take a brief look at both types in the next two sections.

5.1 Problems

As has become clear in the example around Expression (11), solving optimization problems is an important part of working with the models we propose. We have not given a full account of the solution to the optimization problem in that example. One reason is that the focus of this paper is on the description of the conjugate and predictive models. Another reason is that we judge the lengthy description would be of little added value.

Devising approximation algorithms to solve the optimization problem might be necessary. The problem with these is that, because we want conservative inferences, an outer approximation has to be used. For example, if in Criterion (10), we would be satisfied with a local minimum

instead of the global minimum over $\mathcal{Y}_C$ (an inner approximation), the resulting '$>$'-relation would be too strong. This means that a class could be preferred to another even though it is not warranted by the model. This implies that there could be maximal elements of the resulting partial order that would not have been maximal if the global minimum had been used.

5.2 Prospects for further research

We too can include the standard disclaimer about ample prospects for further research. For one thing, one could investigate other exponential families, whose probability functions have a more general form than the one given by Equation (1).[9]

Closer to the focus of this paper, we could investigate if it is possible to find lower and upper cumulative distribution functions for (some) predictive models, as has already been done for the IDMM [13].

In the paper introducing the naive credal classifier [14], a section is devoted to the case of coping with missing data. The approach taken there can theoretically also be used when the attribute values are continuous. Letting a missing attribute value a correspond to a subset of $\mathcal{A}$, we just add one more optimization problem over this subset. How this works out in practice remains to be investigated.

The issue of missing data in the training set of a credal classifier, or more generally, of noisy samples, will have its effect on the updating process and thus on the form of the sets $\mathcal{Y}_{\mathcal{A}|C}$, and is also interesting to look at.

A last issue, for all conjugate and predictive models, is *forgetting*. In some applications it might be interesting to not let the model become too precise. This can be achieved in an *ad-hoc* way by manipulating the number of counts n outside of the updating process. It would be interesting to look for approaches that are better justified.

Acknowledgments

This paper presents research results of the Belgian Program on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office. The scientific responsibility rests with its authors.

Erik Quaeghebeur's research is financed by a Ph.D. grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen).

The authors wish to thank the reviewers for helpful suggestions.

[9] To be specific, following the nomenclature in the literature [7, 1], we could look at *nonregular, curved or noncanonical exponential families*.

References

[1] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley, 1994.

[2] Agata Boratyńska. Stability of Bayesian inference in exponential families. *Stat. Probabil. Lett.*, 36:173–178, 1997.

[3] Remco R. Bouckaert. Naive Bayes Classifiers that Perform Well with Continuous Variables. In G. I. Webb and X. Yu, editors, *AI 2004: Advances in Artificial Intelligence: 17th Australian Joint Conference on Artificial Intelligence*, pages 1089–1094, 2004.

[4] F. P. A. Coolen. Imprecise conjugate prior densities for the one-parameter exponential family of distributions. *Stat. Probabil. Lett.*, 16:337–342, 1993.

[5] Persi Diaconis and Donald Ylvisaker. Conjugate priors for exponential families. *Ann. Stat.*, 7:269–281, 1979.

[6] George H. John and Pat Langley. Estimating Continuous Distributions in Bayesian Classifiers. In Ph. Besnard and S. Hanks, editors, *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 338–345, 1995.

[7] Samuel Kotz, N. Balakrishnan, and Norman L. Johnson. *Continuous Multivariate Distributions*, volume 1. Wiley, 2nd edition, 2000.

[8] Teddy Seidenfeld and Larry Wasserman. Dilation for sets of probabilities. *Ann. Stat.*, 21:1139–1154, 1993.

[9] Matthias C. M. Troffaes. Decision Making with Imprecise Probabilities: A Short Review. The SIPTA newsletter, December 2004.

[10] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.

[11] Peter Walley. Inferences from Multinomial Data: Learning about a Bag of Marbles. *J. Roy. Stat. Soc. B Met.*, 58:3–57, 1996.

[12] Peter Walley. A Bounded Derivative Model for Prior Ignorance about a Real-valued Parameter. *Scand. J. Stat.*, 24:463–483, 1997.

[13] Peter Walley and Jean-Marc Bernard. Imprecise Probabilistic Prediction for Categorical Data. Technical Report CAF-9901, Université de Paris 8, 1999.

[14] Marco Zaffalon. Statistical inference of the naive credal classifier. In G. de Cooman, T. L. Fine, and T. Seidenfeld, editors, *ISIPTA '01: Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, pages 384–393, 2001.