

Imprecise probability models for inference in exponential families

Erik Quaeghebeur & Gert de Cooman

SYSTeMS research group



Who we are & what we do

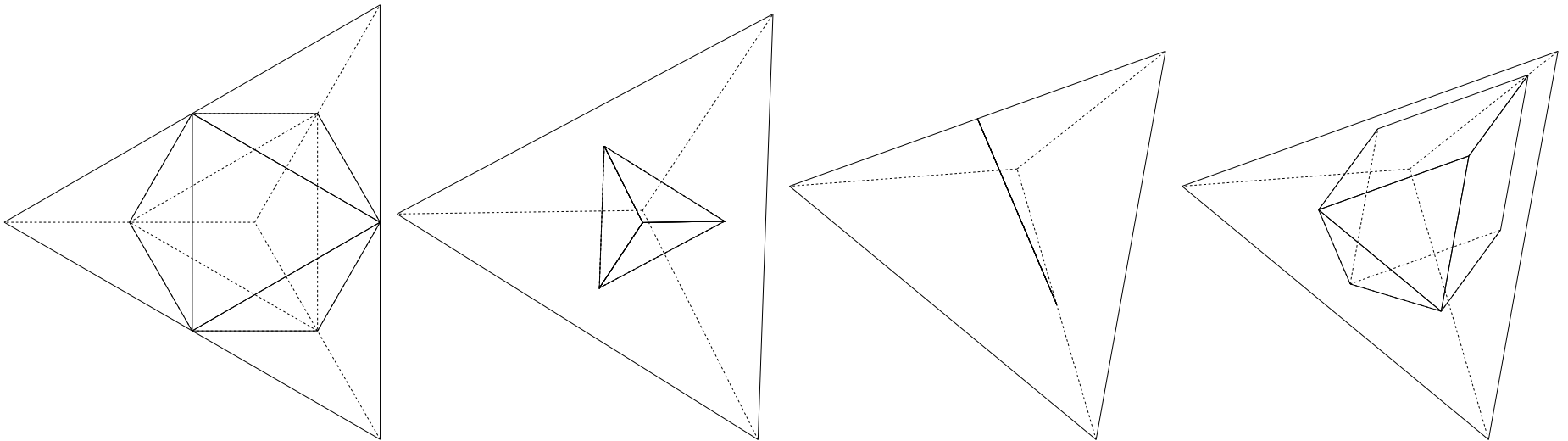
- *Erik Quaeghebeur*, PhD student of *Gert de Cooman*, SYSTeMS research group, Ghent University, Belgium.

Who we are & what we do

- *Erik Quaeghebeur*, PhD student of *Gert de Cooman*, SYSTeMS research group, Ghent University, Belgium.
- Current research interests:

Who we are & what we do

- *Erik Quaeghebeur*, PhD student of *Gert de Cooman*, SYSTeMS research group, Ghent University, Belgium.
- Current research interests:
 - *extreme lower probabilities;*

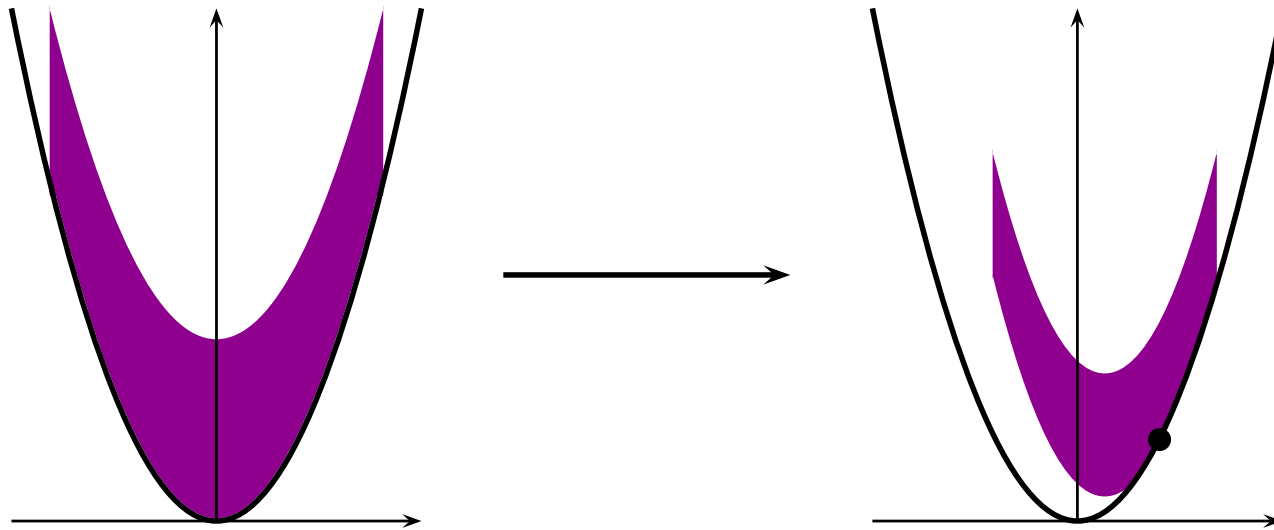


Who we are & what we do

- *Erik Quaeghebeur*, PhD student of *Gert de Cooman*, SYSTeMS research group, Ghent University, Belgium.
- Current research interests:
 - extreme lower probabilities;
 - *(partition) exchangeability*;

Who we are & what we do

- Erik Quaeghebeur, PhD student of Gert de Cooman, SYSTeMS research group, Ghent University, Belgium.
- Current research interests:
 - extreme lower probabilities;
 - (partition) exchangeability;
 - *exponential families*.



Socratic dialogue

Our poster: the technical details

- Theory on the left. . .
- . . . examples on the right.
- Let me guide you through.
- Ask me questions.

**IMPRECISE PROBABILITY MODELS
FOR INFERENCE IN EXPONENTIAL FAMILIES**
 ERIK QUAEGBEUR & GERT DE COOMAN
 SYSTeMS Research Group
 Department of Electrical Energy, Systems & Automation, Ghent University
 Technologiepark 914, B-9002 Zwijnaarde, Belgium
 {Erik.Quaeghebeur, Gert.deCooman}@UGent.be

EXPONENTIAL FAMILIES

An exponential family
Consider taking i.i.d. samples x (sample space X) of a random variable that is distributed according to an exponential family with probability function of the form

$$E(x|\theta) = a(x) \exp(\theta \cdot T(x) - b(\theta)),$$

with functions $a : X \rightarrow \mathbb{R}^+$, $b : \Psi \rightarrow \mathbb{R}$ and with canonical parameter $\theta \in \Psi$ and sufficient statistic $T : X \rightarrow \mathbb{R}^+$.

The conjugate family
By looking at $E(x|\theta)$ as a likelihood function $L_\theta : \Psi \rightarrow \mathbb{R}^+$, we can write down the probability density function of the corresponding family of conjugate distributions,

$$CE(x|\mu, \nu) = c(\mu, \nu) \exp(\mu \cdot T(x) - b(\mu)),$$

with normalization factor c and two parameters which can be given specific interpretations: μ (pseudocount) $\mu \in \mathbb{R}^+$ and an average sufficient statistic $\nu \in Y \subset \mathcal{Y}$.

The predictive family
The probability function of the corresponding family of predictive distributions can be derived by combining L_θ and $CE(x|\mu, \nu)$.

$$PE(x|\mu, \nu) = \int_{\Psi} CE(x|\mu, \nu) L_\theta = \frac{c(\mu, \nu) a(x)}{c(\mu, \nu) + 1} \frac{E(x|\theta)}{E(x|\theta)}$$

Example: Multinomial sampling
In this case, the one sample likelihood function is a multivariate Bernoulli $B(x|\theta)$, the conjugate density function is a Dirichlet $Dir(\mu, \nu)$ and the predictive mass function is a Dirichlet multinomial $DM(x|\mu, \nu)$, where

$$x \in \{0, 1\}^d : \|x\|_1 \leq 1; \theta \in (0, 1]^d : \|\theta\|_1 = 1, \theta_0 = 1 - \sum_{i=1}^d \theta_i; T(x) = x; \theta(\theta) = \left(\frac{\Gamma(\sum_{i=1}^d \theta_i)}{\prod_{i=1}^d \Gamma(\theta_i)} \right);$$

$$y \in \{0, 1\}^d : \|y\|_1 < 1, y_0 = 1 - \sum_{i=1}^d y_i; a = 1; b(\theta(\theta)) = \ln(\theta(\theta)); c(\mu, \nu) = \frac{\Gamma(\mu)}{\prod_{i=1}^d \Gamma(\mu_i) \Gamma(\nu)}$$

Example: Normal sampling
Now, the one sample likelihood function is a Normal $N(x|\mu, \sigma^2)$, the conjugate density function is a Normal-gamma

$$Ng(\mu|\gamma, \alpha) \propto \Gamma(\alpha) \exp\left(-\frac{\alpha-1}{2} \frac{\mu^2 + \gamma^2}{\alpha-1}\right)$$

and the predictive density function is a Student $St(x|\gamma, \alpha)$,

$$St(x|\gamma, \alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha-1)} \frac{\gamma}{\sqrt{\alpha-1}} \left(1 + \frac{\gamma^2 + x^2}{\alpha-1}\right)^{-\alpha}$$

where $\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+, \alpha \in \mathbb{R}^+, \gamma \in \mathbb{R}^+$; $\theta(x) = (x, \sigma^2)$; $a(\theta(x)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$; $b(\theta(x)) = \frac{1}{2\sigma^2} \ln(\sigma^2)$; $c(\mu, \nu) = \frac{\Gamma(\alpha)}{\Gamma(\alpha-1)}$

The conjugate model
The conjugate model for inference in an exponential family is a lower prevision, defined as the lower envelope of a set of linear previsions that correspond to members of the conjugate family:

$$\underline{P}_\theta(f|\mu^0, \nu^0) = \inf_{\theta \in \Psi} P_\theta(f|\mu^0, \nu^0), \text{ where } P_\theta(f|\mu^0, \nu^0) = \int CE(f|\mu^0, \nu^0) L_\theta, f \in \mathcal{L}(\Psi).$$

Here, $\mathcal{L}(\Psi)$ is the set of all measurable gambles (bounded functions) on Ψ and Ψ^0 is some subset of Ψ .

The predictive model
The predictive model for inference in an exponential family is defined similarly:

$$\underline{P}_\theta(f|\mu^0, \nu^0) = \inf_{\theta \in \Psi} P_\theta(f|\mu^0, \nu^0), \text{ where } P_\theta(f|\mu^0, \nu^0) = \int PE(f|\mu^0, \nu^0) L_\theta, f \in \mathcal{L}(X).$$

A prior choice μ^0 and bounded subset Ψ^0 of Ψ for the parameters of these models must be made. When k samples are taken—with sufficient statistic $T^k \in \mathcal{Y}^k$ —, these can be used to update the models (Bayes' rule) by obtaining posterior parameters

$$\mu^k = \mu^0 + k, \quad \nu^k = \frac{\mu^0 \nu^0 + T^k}{k+1}, \quad y \in \Psi^k \subset \mathcal{Y}^k.$$

The imprecision of the inferences of these models are proportional to the volume of $\text{co}(\Psi^k)$. So the imprecision decreases with k at a rate that decreases with μ^0 .

Example of updating: Multinomial sampling

Example of updating: Normal sampling

AN APPLICATION: CLASSIFICATION

Credal classification
A classifier maps attribute values $a \in \mathcal{A}$ to one or more classes $c \in C$. In a credal classifier, a conditional lower prevision $\underline{D}_c | \mathcal{A}$ on $\mathcal{L}(C)$ is used to make pairwise comparisons of classes c^i and c^j , given attribute values a . The criterion used is

$$c^i > c^j \Leftrightarrow \underline{D}_c(c^i | a) - \underline{D}_c(c^j | a) > 0.$$

The maximal elements of the resulting partial order are the output of the classifier. The computational complexity of the optimization problem that has to be solved for comparing two classes c^i and c^j depends highly on the type of attributes that are used.

Creating a credal classifier
We derive $\underline{D}_c | \mathcal{A}$ by conditioning a joint lower prevision \underline{D} on $\mathcal{L}(C \times \mathcal{A})$. \underline{D} is the marginal extension of a class model \underline{D}_C on $\mathcal{L}(C)$ and an attribute model $\underline{D}_\mathcal{A}$ on $\mathcal{L}(\mathcal{A})$. When the number of classes is finite and the attribute values are distributed according to an exponential family, we can use predictive models $\underline{D}_C(c^i | \mathcal{A})$ and $\underline{D}_\mathcal{A}(y | \mathcal{A})$ for the class and attribute models.

Example optimization problem: multiple discrete attributes

$$c^i > c^j \Leftrightarrow \inf_{\theta \in \Psi} \left[\sum_{k=1}^K \inf_{a \in \mathcal{A}^k} \left(\sum_{c^i \in C} \theta_{c^i} \underline{D}_C(c^i | a) - \sum_{c^j \in C} \theta_{c^j} \underline{D}_C(c^j | a) \right) \right] > 0$$

The \inf / \sup over Ψ of θ_{c^i} are simple functions of ν , that guarantee the convexity of the objective function. So this problem can easily be solved numerically.

Example optimization problem: one normal attribute
The criterion is the same as above, but with the sums replaced by the \inf / \sup over Ψ of

$$\frac{\theta_{c^i}}{\sqrt{\theta_{c^i} + 1}} \frac{\Gamma(\frac{\theta_{c^i} + 1}{2})}{\Gamma(\frac{\theta_{c^i}}{2})} \left[\theta_{c^i} \nu_{c^i, 2} - \nu_{c^i, 1} \right] - \frac{\theta_{c^j}}{\sqrt{\theta_{c^j} + 1}} \frac{\Gamma(\frac{\theta_{c^j} + 1}{2})}{\Gamma(\frac{\theta_{c^j}}{2})} \left[\theta_{c^j} \nu_{c^j, 2} - \nu_{c^j, 1} \right]$$

It is not yet clear if and how this problem can be solved.

- p.4/4

