

# IMPRECISE PROBABILITY MODELS FOR INFERENCE IN EXPONENTIAL FAMILIES

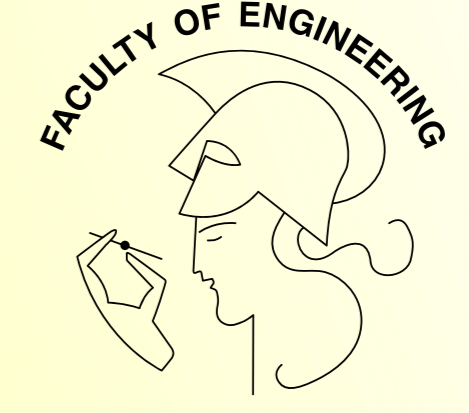
ERIK QAEGHEBEUR & GERT DE COOMAN

SYSTeMS Research Group

Department of Electrical Energy, Systems & Automation, Ghent University

Technologiepark 914, B-9052 Zwijnaarde, Belgium

{Erik.Quaeghebeur, Gert.deCooman}@UGent.be



## EXPONENTIAL FAMILIES

### An exponential family

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

$$\text{Ef}(x|\psi) = a(x) \exp(\langle \psi, \tau(x) \rangle - b(\psi)),$$

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

### The conjugate family

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

$$\text{CEf}(\psi|n, y) = c(n, y) \exp(n[\langle \psi, y \rangle - b(\psi)]),$$

with normalization factor  $c$  and two parameters which can be given specific interpretations:  $a$  (pseudo)count  $n \in \mathbb{R}^+$  and an average sufficient statistic  $y \in \mathcal{Y} = \text{co}(\mathcal{T})$ .

### The predictive family

The probability function of the corresponding family of predictive distributions can be derived by combining  $L_x$  and  $\text{CEf}(\cdot|n, y)$ ,

$$\text{PEf}(x|n, y) = \int_{\Psi} \text{CEf}(\cdot|n, y) L_x = \frac{c(n, y) a(x)}{c(n+1, \frac{ny+\tau(x)}{n+1}}.$$

### Example: Multinomial sampling

In this case, the one sample likelihood function is a multivariate Bernoulli  $\text{Br}(x|\theta)$ , the conjugate density function is a Dirichlet  $\text{Di}(\theta|ny, ny_0)$  and the predictive mass function is a Dirichlet-multinomial  $\text{DiMn}(x|ny, ny_0)$ , where

$$x \in \{0, 1\}^d: \|x\| \leq 1; \quad \theta \in (0, 1)^d: \|\theta\| < 1, \theta_0 = 1 - \sum_i \theta_i; \quad \tau(x) = x; \quad \psi(\theta) = \left( \ln \frac{\theta_i}{\theta_0} \right)_{i=1}^d;$$

$$y \in (0, 1)^d: \|y\| < 1, y_0 = 1 - \sum_i y_i; \quad a = 1; \quad b(\psi(\theta)) = \ln(\theta_0); \quad c(n, y) = \frac{\Gamma(n)}{\Gamma(ny_0) \prod_i \Gamma(ny_i)}.$$

### Example: Normal sampling

Now, the one sample likelihood function is a Normal  $\text{N}(x|\mu, \lambda)$ , the conjugate density function is a Normal-gamma

$$\text{N}(\mu|y_1, n\lambda) \text{Ga}(\lambda | \frac{n+3}{2}, \frac{n[y_2 - y_1^2]}{2}),$$

and the predictive density function is a Student  $\text{St}(x|y_1, \frac{n+3}{n+1} \frac{1}{y_2 - y_1^2}, n+3)$ , where

$$x \in \mathbb{R}; \quad \mu \in \mathbb{R}, \lambda \in \mathbb{R}^+, \sigma^2 = \frac{1}{\lambda}; \quad \tau(x) = (x, x^2); \quad \psi(\lambda, \mu) = (\lambda\mu, -\frac{1}{2}\lambda);$$

$$y \in \mathbb{R} \times \mathbb{R}^+: y_2 - y_1^2 > 0; \quad a = \frac{1}{\sqrt{2\pi}}; \quad b(\psi(\mu, \lambda)) = \frac{\lambda\mu^2 - \ln(\lambda)}{2}; \quad c(n, y) = \frac{2\sqrt{n}}{\sqrt{2\pi}} \frac{\left[ \frac{n[y_2 - y_1^2]}{2} \right]^{\frac{n+3}{2}}}{\Gamma(\frac{n+3}{2})}.$$

## IMPRECISE PROBABILITY MODELS

### 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:

$$\underline{P}_C(f|n^k, \mathcal{Y}^k) = \inf_{y \in \mathcal{Y}^k} P_C(f|n^k, y), \text{ where } P_C(f|n^k, y) = \int_{\Psi} \text{CEf}(\cdot|n^k, y) f, f \in \mathcal{L}(\Psi).$$

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

### The predictive model

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

$$\underline{P}_P(f|n^k, \mathcal{Y}^k) = \inf_{y \in \mathcal{Y}^k} P_P(f|n^k, y), \text{ where } P_P(f|n^k, y) = \int_{\mathcal{X}} \text{PEf}(\cdot|n^k, y) f, f \in \mathcal{L}(\mathcal{X}).$$

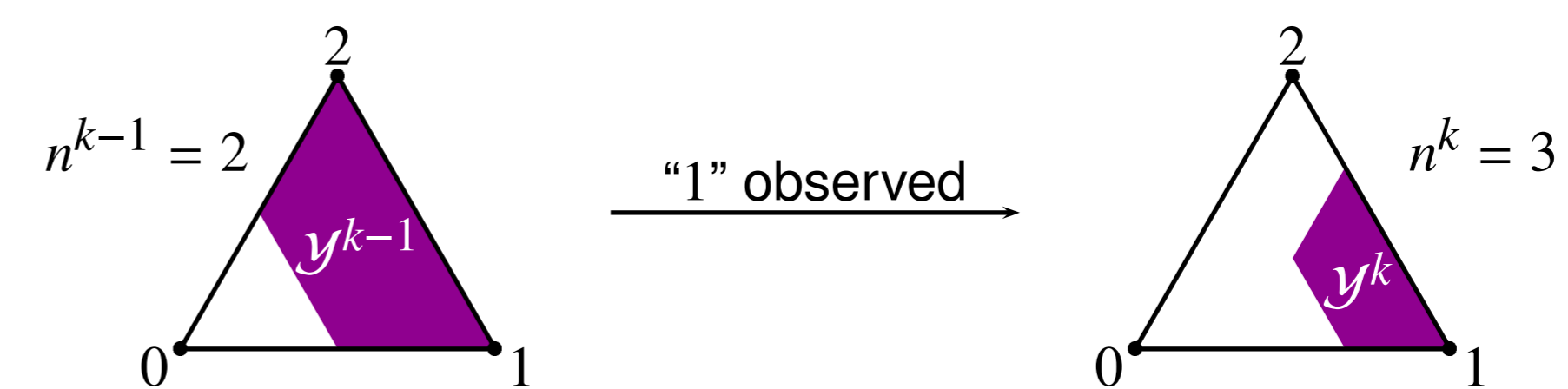
### Updating and imprecision

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

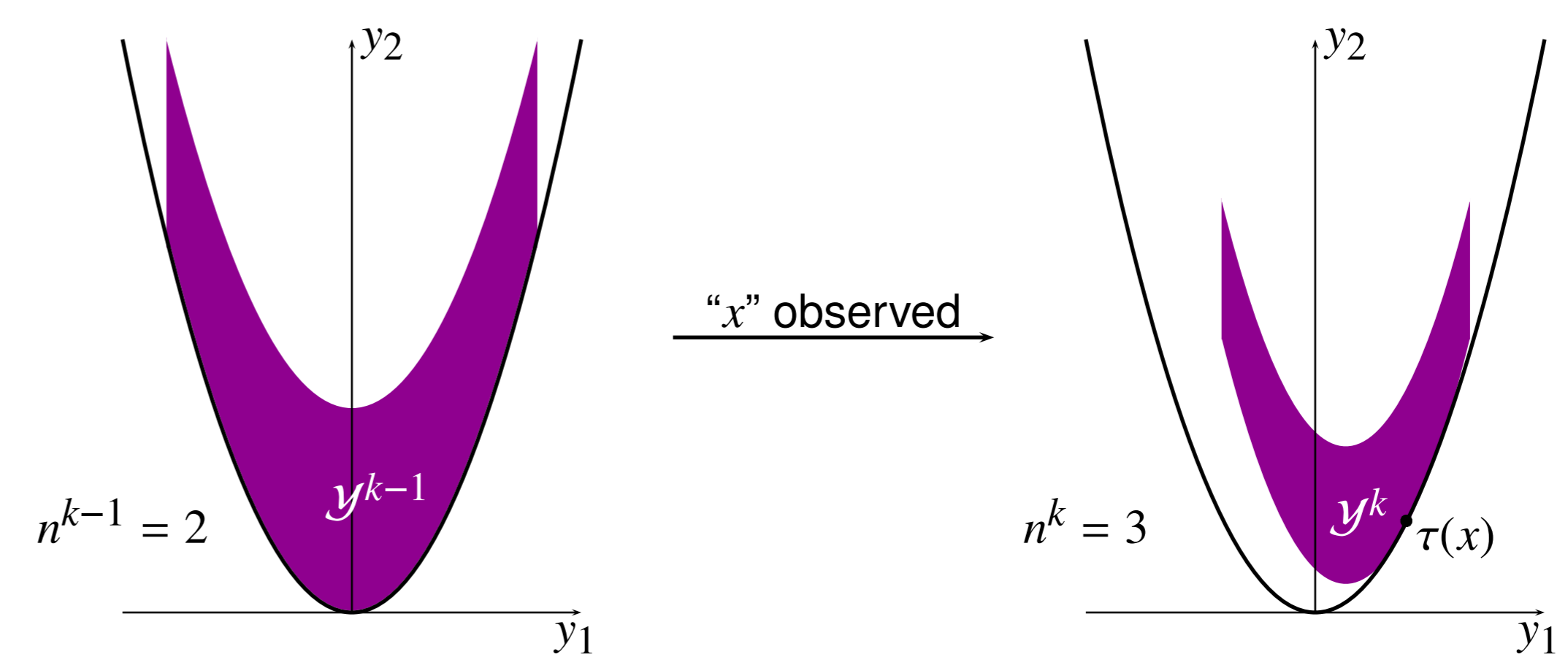
$$n^k = n^0 + k, \quad \mathcal{Y}^k = \left\{ \frac{n^0 y + \tau^k}{n^0 + k} : y \in \mathcal{Y}^0 \right\} \subset \mathcal{Y}.$$

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

### 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  $c \in \mathcal{C}$ . In a credal classifier, a conditional lower prevision  $\underline{P}(\cdot|a)$  on  $\mathcal{L}(\mathcal{C})$  is used to make pairwise comparisons of classes  $c'$  and  $c''$ , given attribute values  $a$ . The criterion used is

$$c' > c'' \Leftrightarrow \underline{P}(I_{c'} - I_{c''} | a) > 0.$$

The maximal elements of the resulting strict partial order are the output of the classifier.

The computational complexity of the optimization problem that has to be solved for comparing two classes  $c'$  and  $c''$  depends highly on the type of attributes that are used.

### Creating a credal classifier

We derive  $\underline{P}(\cdot|a)$  by conditioning a joint lower prevision  $\underline{E}$  on  $\mathcal{L}(\mathcal{C} \times \mathcal{A})$ .  $\underline{E}$  is the marginal extension of a class model  $\underline{P}$  on  $\mathcal{L}(\mathcal{C})$  and an attribute model  $\underline{P}(\cdot|c)$  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  $\underline{P}_P(\cdot|n_C, \mathcal{Y}_C)$  and  $\underline{P}_P(\cdot|n_{\mathcal{A}|c}, \mathcal{Y}_{\mathcal{A}|c})$  for the class and attribute models.

### Example optimization problem: multiple discrete attributes

$$c' > c'' \Leftrightarrow \inf_{y \in \mathcal{Y}_C} \left[ y_{c'} \prod_i \inf_{y_{\mathcal{A}|c'} \in \mathcal{Y}_{\mathcal{A}|c'}} y_{\mathcal{A}|c'} - y_{c''} \prod_i \sup_{y_{\mathcal{A}|c''} \in \mathcal{Y}_{\mathcal{A}|c''}} y_{\mathcal{A}|c''} \right] > 0.$$

The inf / sup  $y_{\mathcal{A}|c} \in \mathcal{Y}_{\mathcal{A}|c}$  of  $y_{\mathcal{A}|c}$  are simple functions of  $y_C$  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 products replaced by the inf / sup  $y_{\mathcal{A}|c} \in \mathcal{Y}_{\mathcal{A}|c}$  of

$$\frac{\sqrt{n_{\mathcal{A}|c}} \Gamma(\frac{n_{\mathcal{A}|c}+4}{2})}{\sqrt{n_{\mathcal{A}|c}+1} \Gamma(\frac{n_{\mathcal{A}|c}+3}{2})} \frac{[n_{\mathcal{A}|c} y_{\mathcal{A}|c,2} - n_{\mathcal{A}|c} y_{\mathcal{A}|c,1}^2]^{\frac{n_{\mathcal{A}|c}+3}{2}}}{[n_{\mathcal{A}|c} y_{\mathcal{A}|c,2} + a^2 - \frac{1}{n_{\mathcal{A}|c}+1} [n_{\mathcal{A}|c} y_{\mathcal{A}|c,1} + a]^2]^{\frac{n_{\mathcal{A}|c}+4}{2}}}.$$

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