

# Imprecise probability models for inference in exponential families

*SYSTeMS-dialogue of 14 July 2005*

Erik Quaeghebeur

SYSTeMS research group



# Overview

1. The general idea
2. Specifying the details
3. A useful result
4. Updating
5. History: how this research got started
6. An application: classification
7. Conclusions

# The general idea

# The general idea

- *Sampling model*  $f(x | \psi)$ : likelihood function  $L_x(\psi)$ , sufficient statistic.

# The general idea

- *Sampling model*  $f(x | \psi)$ : likelihood function  $L_x(\psi)$ , sufficient statistic.
- Choose some *prior*  $C(\psi)$ : obtain a *posterior* after observing samples.

# The general idea

- *Sampling model*  $f(x | \psi)$ : likelihood function  $L_x(\psi)$ , sufficient statistic.
- Choose some *prior*  $C(\psi)$ : obtain a *posterior* after observing samples.
- Obtain the corresponding *predictive distribution* through 
$$P(x) = \int_{\Psi} CL_x.$$

# The general idea

- *Sampling model*  $f(x | \psi)$ : likelihood function  $L_x(\psi)$ , sufficient statistic.
- Choose some *prior*  $C(\psi)$ : obtain a *posterior* after observing samples.
- Obtain the corresponding *predictive distribution* through  $P(x) = \int_{\Psi} CL_x$ .
- Obtain the corresponding *linear previsions*  $P_C$  and  $P_P$ .

# The general idea

- *Sampling model*  $f(x | \psi)$ : likelihood function  $L_x(\psi)$ , sufficient statistic.
- Choose some *prior*  $C(\psi)$ : obtain a *posterior* after observing samples.
- Obtain the corresponding *predictive distribution* through  $P(x) = \int_{\Psi} CL_x$ .
- Obtain the corresponding *linear previsions*  $P_C$  and  $P_P$ .
- Imprecision: take a set of priors, use the lower envelope theorem to obtain coherent *lower previsions*  $\underline{P}_C$  and  $\underline{P}_P$ .

# Specifying the details: sampling model

- *Sampling model*  $f(x | \psi)$ : likelihood function  $L_x(\psi)$ , sufficient statistic.

# Specifying the details: sampling model

- *Exponential family sampling model*  $Ef(x | \psi)$ : likelihood  $L_x(\psi)$ , sufficient statistic  $\tau(x)$  of fixed dimension.

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

# Specifying the details: sampling model

- *Exponential family sampling model*  $Ef(x | \psi)$ : likelihood  $L_x(\psi)$ , sufficient statistic  $\tau(x)$  of fixed dimension.

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

**Multinomial sampling** Likelihood function is a *multivariate Bernoulli*  $Br(x | \theta)$ :

$$x \in \{0, 1\}^d : \|x\| \leq 1; \quad \tau(x) = x;$$

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

$$a = 1; \quad b(\psi(\theta)) = \ln(\theta_0).$$

# Specifying the details: sampling model

- *Exponential family sampling model*  $\text{Ef}(x | \psi)$ : likelihood  $L_x(\psi)$ , sufficient statistic  $\tau(x)$  of fixed dimension.

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

**Normal sampling** Likelihood is a *Normal*  $N(x | \mu, \lambda)$ :

$$x \in \mathbb{R}; \quad \tau(x) = (x, x^2);$$

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

$$a = \frac{1}{\sqrt{2\pi}}; \quad b(\psi(\mu, \lambda)) = \frac{\lambda\mu^2 - \ln(\lambda)}{2}.$$

# Specifying the details: conjugate

- Choose some *prior*  $C(\psi)$ : obtain a *posterior* after observing samples.

# Specifying the details: conjugate

- Choose a *conjugate prior*  $\text{CEf}(\psi | n^0, y^0)$ : easily obtain a *posterior*  $\text{CEf}(\psi | n^k, y^k)$  after observing  $k$  samples.

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

# Specifying the details: conjugate

- Choose a *conjugate prior*  $\text{CEf}(\psi | n^0, y^0)$ : easily obtain a *posterior*  $\text{CEf}(\psi | n^k, y^k)$  after observing  $k$  samples.

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

**Multinomial sampling** The conjugate distribution is a *Dirichlet distribution*  $\text{Di}(\theta | ny, ny_0)$ :

$$y \in (0, 1)^d : \|y\| < 1, y_0 = 1 - \sum_i y_i;$$

$$c(n, y) = \frac{\Gamma(n)}{\Gamma(ny_0) \prod_i \Gamma(ny_i)}.$$

# Specifying the details: conjugate

- Choose a *conjugate prior*  $\text{CEf}(\psi \mid n^0, y^0)$ : easily obtain a *posterior*  $\text{CEf}(\psi \mid n^k, y^k)$  after observing  $k$  samples.

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

**Normal sampling** The conjugate distribution is a *Normal-gamma distribution*

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

$$y \in \mathbb{R} \times \mathbb{R}^+ : y_2 - y_1^2 > 0;$$

$$c(n, y) = \frac{2\sqrt{n}}{\sqrt{2\pi}} \frac{\left[ \frac{n[y_2 - y_1^2]}{2} \right]^{\frac{n+3}{2}}}{\Gamma(\frac{n+3}{2})}.$$

# Specifying the details: predictive

- Obtain the corresponding *predictive distribution* through  
$$P(x) = \int_{\Psi} CL_x.$$

# Specifying the details: predictive

- Obtain the corresponding *predictive distribution* through

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

# Specifying the details: predictive

- Obtain the corresponding *predictive distribution* through

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

**Multinomial sampling** The predictive distribution is a *Dirichlet-multinomial distribution*  $\text{DiMn}(x | ny, ny_0)$ .

**Normal sampling** The predictive distribution is a *Student distribution*  $\text{St}(x | y_1, \frac{n+3}{n+1} \frac{1}{y_2 - y_1^2}, n + 3)$ .

# Specifying the details: linear previsions

- Obtain the corresponding *linear previsions*  $P_C$  and  $P_P$ .

# Specifying the details: linear previsions

- Obtain the corresponding *linear previsions*

$$P_C(f \mid n^k, y) = \int_{\Psi} \text{CEf}(\cdot \mid n^k, y) f, \quad f \in \mathcal{L}(\Psi) \approx [\Psi \rightarrow \mathbb{R}].$$

and

$$P_P(f \mid n^k, y) = \int_{\mathcal{X}} \text{PEf}(\cdot \mid n^k, y) f, \quad f \in \mathcal{L}(\mathcal{X}) \approx [\mathcal{X} \rightarrow \mathbb{R}].$$

# Specifying the details: lower previsions

- Imprecision: take a set of priors, use the lower envelope theorem to obtain coherent *lower previsions*  $\underline{P}_C$  and  $\underline{P}_P$ .

# Specifying the details: lower previsions

- Imprecision: take a set of priors, one for every  $y \in \mathcal{Y}^0$ , use the lower envelope theorem to obtain coherent *lower previsions*

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

and

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

# A useful result

$$P(\tau | \psi) = \int_{\mathcal{X}} \text{Ef}(\cdot | \psi) \tau$$

# A useful result

$$P(\tau | \psi) = \int_{\mathcal{X}} \text{Ef}(\cdot | \psi) \tau$$

**Multinomial sampling**  $P(\tau | \psi) = \theta(\psi)$ .

**Normal sampling**  $P(\tau | \psi) = (\mu(\psi), m_2(\psi))$ .

# A useful result

$$P(\tau | \Psi)$$

# A useful result

$$P_{\mathcal{C}}(P(\tau | \Psi) | n^k, y^k) = y^k$$

# A useful result

$$\underline{P}_C(P(\tau | \Psi) | n^k, \mathcal{Y}^k) = \inf \mathcal{Y}^k$$

# A useful result

$$\underline{P}_C(P(\tau | \Psi) | n^k, \mathcal{Y}^k) = \inf \mathcal{Y}^k$$

$$\overline{P}_C(P(\tau | \Psi) | n^k, \mathcal{Y}^k) = \sup \mathcal{Y}^k$$

# Updating

- Initial choice  $n^0 \in \mathbb{R}^+$  and  $\mathcal{Y}^0 \subset \mathcal{Y}$  (bounded).

# Updating

- Initial choice  $n^0 \in \mathbb{R}^+$  and  $\mathcal{Y}^0 \subset \mathcal{Y}$  (bounded).
- Take  $k$  samples, keep sufficient statistic  $\tau^k$ .

# Updating

- Initial choice  $n^0 \in \mathbb{R}^+$  and  $\mathcal{Y}^0 \subset \mathcal{Y}$  (bounded).
- Take  $k$  samples, keep sufficient statistic  $\tau^k$ .
- Update parameters:

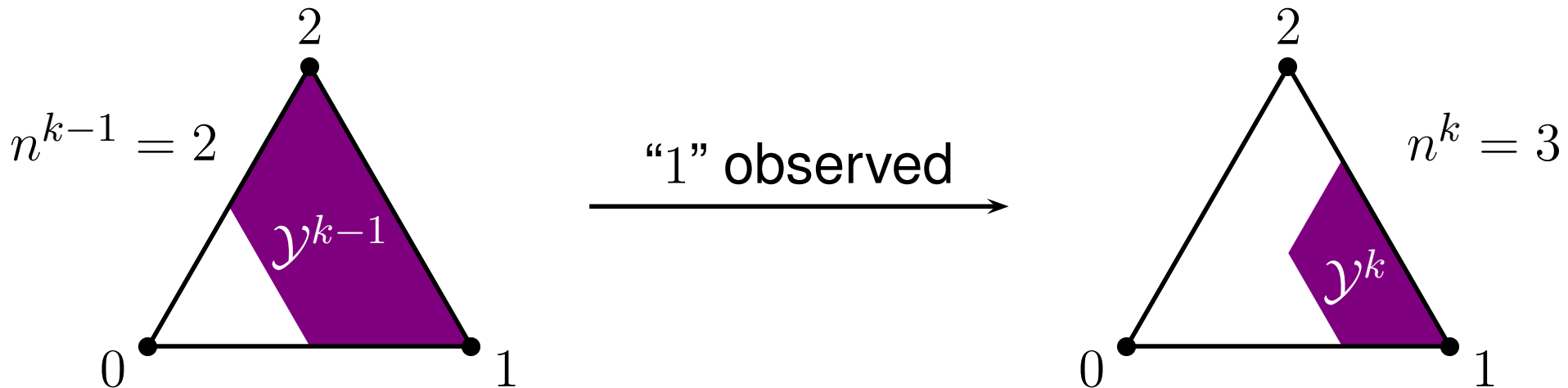
$$n^k = n^0 + k, \quad \mathcal{Y}^k = \left\{ \frac{n^0 y + \tau^k}{n^0 + k} : y \in \mathcal{Y}^0 \right\} \subset \mathcal{Y}.$$

# Updating

- Initial choice  $n^0 \in \mathbb{R}^+$  and  $\mathcal{Y}^0 \subset \mathcal{Y}$  (bounded).
- Take  $k$  samples, keep sufficient statistic  $\tau^k$ .
- Update parameters:

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

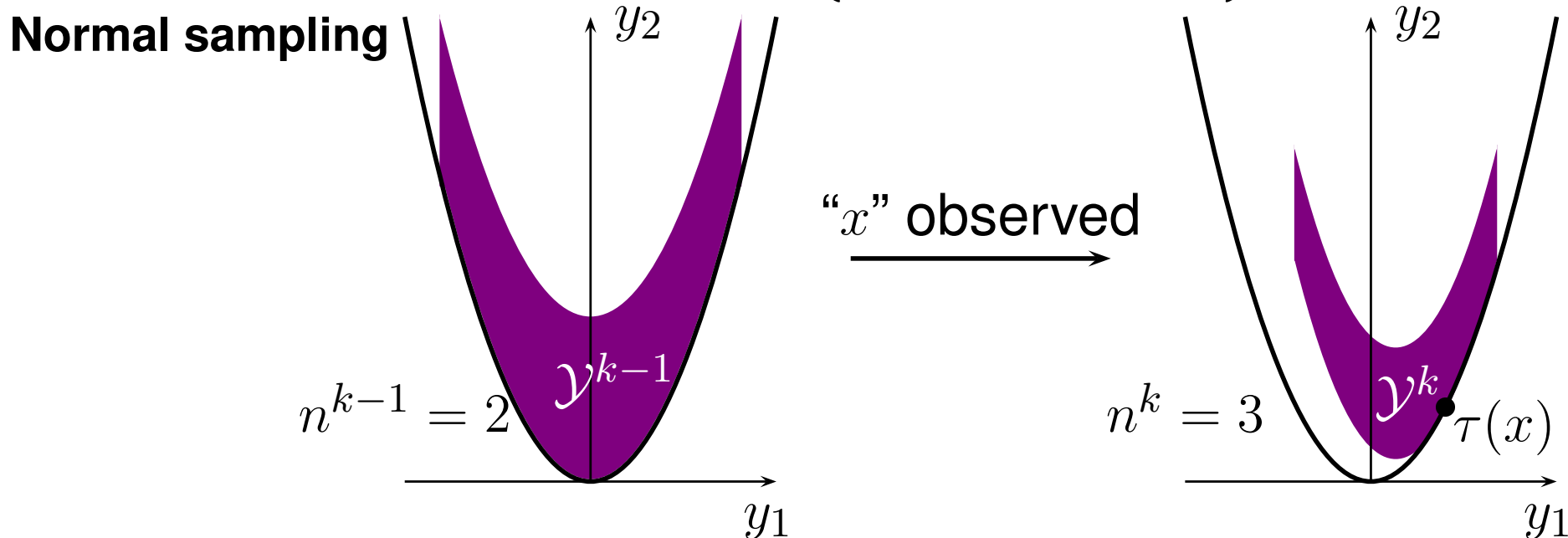
## Multinomial sampling



# Updating

- Initial choice  $n^0 \in \mathbb{R}^+$  and  $\mathcal{Y}^0 \subset \mathcal{Y}$  (bounded).
- Take  $k$  samples, keep sufficient statistic  $\tau^k$ .
- Update parameters:

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



# History: how this research got started

# History: how this research got started

- Using the IDM: problems with optimization problems.
- Literature search: no solution, but...

# History: how this research got started

- Using the IDM: problems with optimization problems.
- Literature search: no solution, but...
- ... the realization that the idea underlying the IDM for multinomial sampling generalizes to all exponential family sampling models:
  - common interpretation for parameters  $n$  and  $y$ ;
  - easy updating.

# History: how this research got started

- Using the IDM: problems with optimization problems.
- Literature search: no solution, but...
- ... the realization that the idea underlying the IDM for multinomial sampling generalizes to all exponential family sampling models:
  - common interpretation for parameters  $n$  and  $y$ ;
  - easy updating.
- However, using these models: again possible problems with optimization problems.

# An application: classification

- Classifier: maps attributes  $a \in \mathcal{A}$  to classes  $c \in \mathcal{C}$ .

# An application: classification

- Classifier: maps attributes  $a \in \mathcal{A}$  to classes  $c \in \mathcal{C}$ .
- *Credal classifier*: uses  $\underline{P}(\cdot \mid \mathcal{A})$  on  $\mathcal{L}(\mathcal{C})$  to create an ordering of the classes, i.e.,  $\underline{P}(f_{c'} - f_{c''} \mid a) > 0$ ?

# An application: classification

- Classifier: maps attributes  $a \in \mathcal{A}$  to classes  $c \in \mathcal{C}$ .
- *Credal classifier*: uses  $\underline{P}(\cdot | \mathcal{A})$  on  $\mathcal{L}(\mathcal{C})$  to create an ordering of the classes, i.e.,  $\underline{P}(f_{c'} - f_{c''} | a) > 0$ ?
- $\underline{P}(\cdot | \mathcal{A})$  created using a *class model*  $\underline{P}$  on  $\mathcal{L}(\mathcal{C})$  and *attribute models*  $\underline{P}(\cdot | \mathcal{C})$  on  $\mathcal{L}(\mathcal{A})$ .

# An application: classification

- Classifier: maps attributes  $a \in \mathcal{A}$  to classes  $c \in \mathcal{C}$ .
- *Credal classifier*: uses  $\underline{P}(\cdot | \mathcal{A})$  on  $\mathcal{L}(\mathcal{C})$  to create an ordering of the classes, i.e.,  $\underline{P}(f_{c'} - f_{c''} | a) > 0$ ?
- $\underline{P}(\cdot | \mathcal{A})$  created using a *class model*  $\underline{P}$  on  $\mathcal{L}(\mathcal{C})$  and *attribute models*  $\underline{P}(\cdot | \mathcal{C})$  on  $\mathcal{L}(\mathcal{A})$ .
- Classically, both models are IDMM's; here, any  $\underline{P}_{\mathcal{P}}(\cdot | n_{\mathcal{A}|\mathcal{C}}, \mathcal{Y}_{\mathcal{A}|\mathcal{C}})$  is possible.

# An application: classification

- Classifier: maps attributes  $a \in \mathcal{A}$  to classes  $c \in \mathcal{C}$ .
- *Credal classifier*: uses  $\underline{P}(\cdot | \mathcal{A})$  on  $\mathcal{L}(\mathcal{C})$  to create an ordering of the classes, i.e.,  $\underline{P}(f_{c'} - f_{c''} | a) > 0$ ?
- $\underline{P}(\cdot | \mathcal{A})$  created using a *class model*  $\underline{P}$  on  $\mathcal{L}(\mathcal{C})$  and *attribute models*  $\underline{P}(\cdot | \mathcal{C})$  on  $\mathcal{L}(\mathcal{A})$ .
- Classically, both models are IDMM's; here, any  $\underline{P}_P(\cdot | n_{\mathcal{A}|\mathcal{C}}, \mathcal{Y}_{\mathcal{A}|\mathcal{C}})$  is possible.
- Advantages:
  - allows for continuous attributes;
  - straightforward training.
- Disadvantage: optimization problems are harder to solve.

# An application: classification

- Example optimization problems:

**Multinomial sampling** (i.e., multiple discrete attributes)

$$c' \succ c'' \iff$$

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

# An application: classification

- Example optimization problems:

**Multinomial sampling** (i.e., multiple discrete attributes)

$$c' \succ c'' \iff$$

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

**Normal sampling** (i.e., one normal attribute) Replace the products above by the  $\inf / \sup_{y_{\mathcal{A}|c} \in \mathcal{Y}_{\mathcal{A}|c}}$  of

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

# Conclusions

# Conclusions

- We presented two imprecise probability models for inference in exponential families:
  - one for making inferences about the parameter describing the sampling model;
  - the other for making inferences about future samples.

# Conclusions

- We presented two imprecise probability models for inference in exponential families:
  - one for making inferences about the parameter describing the sampling model;
  - the other for making inferences about future samples.
- Applicable for a large range of sampling models...
- ... and thus potentially useful for many applications.

# Conclusions

- We presented two imprecise probability models for inference in exponential families:
  - one for making inferences about the parameter describing the sampling model;
  - the other for making inferences about future samples.
- Applicable for a large range of sampling models...
- ... and thus potentially useful for many applications.
- However, difficult optimization problems might severely limit their use.

