Imprecise probability models for inference in exponential families

SYSTeMS-dialogue of 14 July 2005

Erik Quaeghebeur

SYSTeMS research group

UNIVERSITEIT GENT

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* $\mathsf{f}(x \mid \psi)$: likelihood function $\mathsf{L}_x(\psi)$, sufficient statistic.

The general idea

- *Sampling model* $\mathsf{f}(x \mid \psi)$: likelihood function $\mathsf{L}_x(\psi)$, sufficient statistic.
- Choose some *prior* $\mathsf{C}(\psi)$: obtain a *posterior* after observing samples.

The general idea

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

The general idea

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

The general idea

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

Specifying the details: sampling model

- *Sampling model* $\mathsf{f}(x \mid \psi)$: likelihood function $\mathsf{L}_x(\psi)$, sufficient statistic.

Specifying the details: sampling model

- *Exponential family sampling model* $\mathsf{Ef}(x \mid \psi)$: likelihood $\mathsf{L}_x(\psi)$, sufficient statistic $\tau(x)$ of fixed dimension.

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

Specifying the details: sampling model

- *Exponential family sampling model* $\mathsf{Ef}(x \mid \psi)$: likelihood $\mathsf{L}_x(\psi)$, sufficient statistic $\tau(x)$ of fixed dimension.

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

Multinomial sampling Likelihood function is a *multivariate Bernoulli* $\mathsf{Br}(x \mid \theta)$:

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

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

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

Specifying the details: sampling model

- *Exponential family sampling model* $\mathsf{Ef}(x \mid \psi)$: likelihood $\mathsf{L}_x(\psi)$, sufficient statistic $\tau(x)$ of fixed dimension.

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

Normal sampling Likelihood is a *Normal* $\mathsf{N}(x \mid \mu, \lambda)$:

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

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

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

Specifying the details: conjugate

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

Specifying the details: conjugate

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

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

Specifying the details: conjugate

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

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

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

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

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

Specifying the details: conjugate

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

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

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

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

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

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

Specifying the details: predictive

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

Specifying the details: predictive

- Obtain the corresponding *predictive distribution* through

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

Specifying the details: predictive

- Obtain the corresponding *predictive distribution* through

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

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

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

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*

$$P_{\mathsf{C}}(f \,|\, n^k, y) = \int_{\Psi} \mathsf{CEf}(\cdot \,|\, n^k, y) f, \quad f \in \mathcal{L}(\Psi) \approx [\Psi \to \mathbb{R}].$$

and

$$P_{\mathsf{P}}(f \,|\, n^k, y) = \int_{\mathcal{X}} \mathsf{PEf}(\cdot \,|\, n^k, y) f, \quad f \in \mathcal{L}(\mathcal{X}) \approx [\mathcal{X} \to \mathbb{R}].$$

Specifying the details: lower previsions

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

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*

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

and

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

A useful result

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

A useful result

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

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

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

A useful result

$$P(\tau \mid \Psi)$$

A useful result

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

A useful result

$$\underline{P}_{\mathsf{C}}(P(\tau \mid \Psi) \mid n^k, \mathcal{Y}^k) = \inf \mathcal{Y}^k$$

A useful result

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

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

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 τ^k.

Updating

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

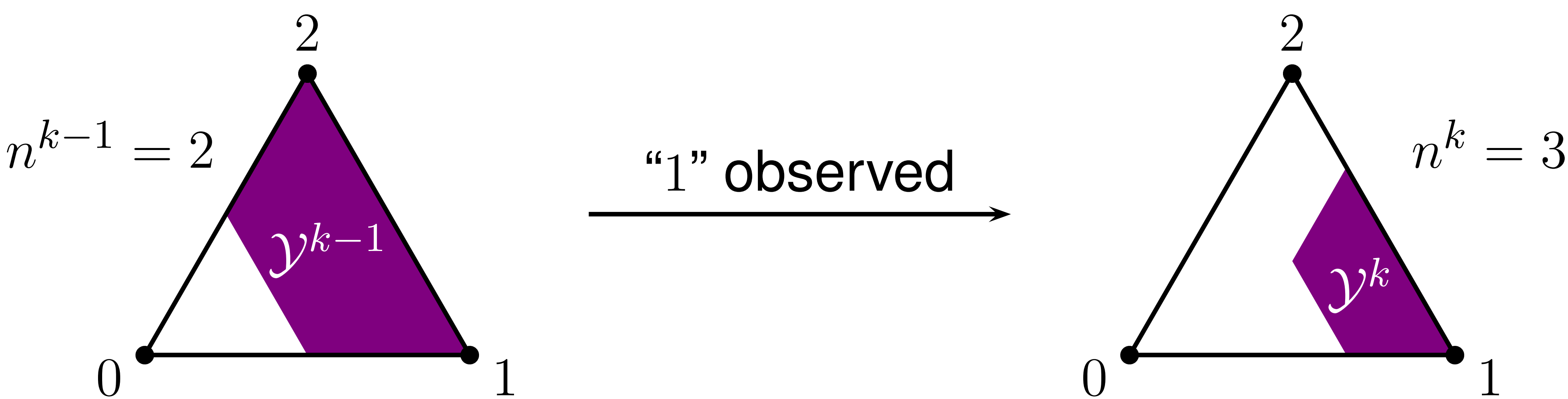
$$n^k = n^0 + k, \qquad \mathcal{Y}^k = \left\{ \frac{n^0 y + \tau^k}{n^0 + k} : y \in \mathcal{Y}^0 \right\} \subset \mathcal{Y}.$$

Updating

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

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

Multinomial sampling

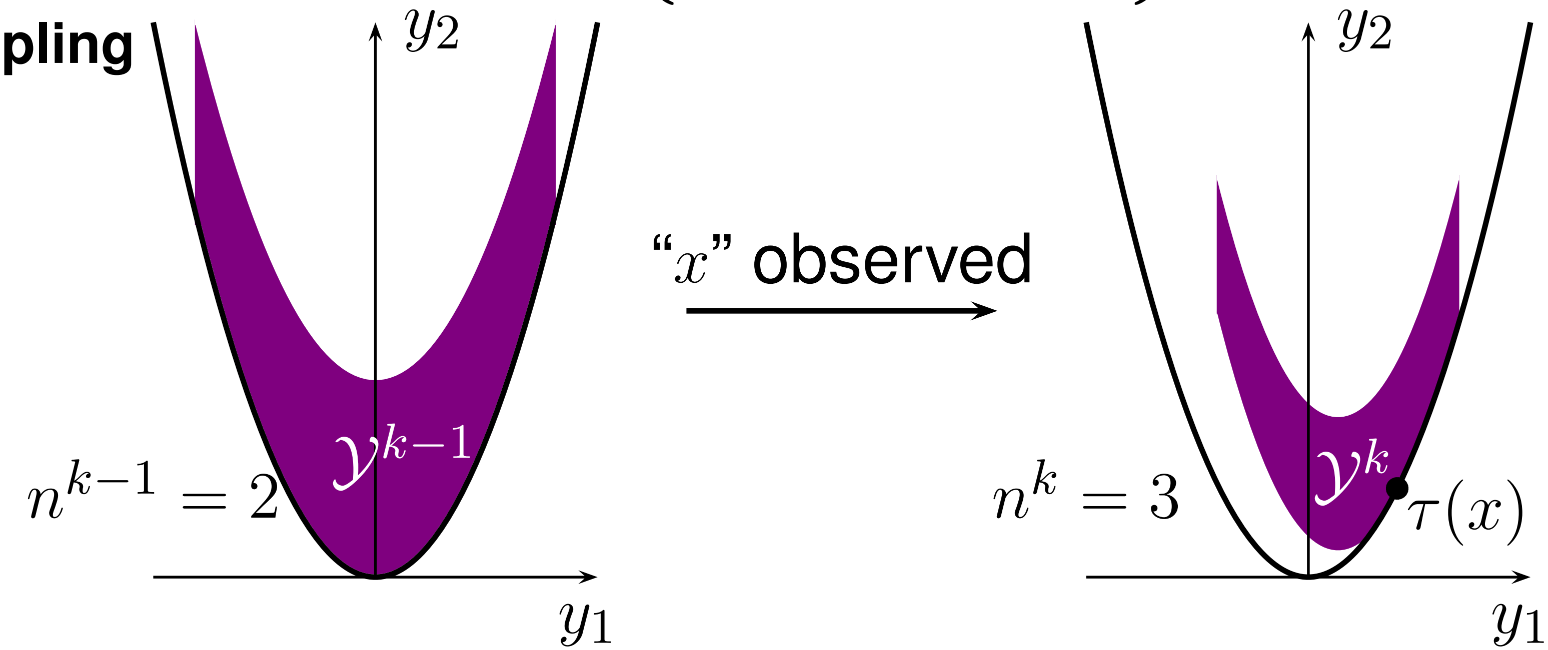


Updating

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

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

Normal sampling



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

An application: classification

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

An application: classification

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

An application: classification

- Example optimizaton 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 optimizaton problems:

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

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

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

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

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

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.

Time for questions!

???????????? ? ????⸮????????