Inferences in hidden Markov models with imprecise probabilities
An efficient algorithm

Gert de Cooman

Ghent University, SYSTeMS
gert.decooman@UGent.be

LMS Lecture
Durham University
26 May 2009
Precise probability models
Mass functions and expectations

Assume we are uncertain about:
- the value or a variable $X$
- in a set of possible values $\mathcal{X}$.

This is usually modelled by a probability mass function $p$ on $\mathcal{X}$:

\[ p(x) \geq 0 \text{ and } \sum_{x \in \mathcal{X}} p(x) = 1; \]

With $p$ we can associate an expectation operator $E_p$:

\[ E_p(f) := \sum_{x \in \mathcal{X}} p(x)f(x) \text{ where } f : \mathcal{X} \rightarrow \mathbb{R}. \]

If $A \subseteq \mathcal{X}$ is an event, then its probability is given by

\[ P_p(A) = \sum_{x \in A} p(x) = E_p(I_A). \]
Precise probability models
The simplex of all probability mass functions

Consider the simplex $\Sigma_\mathcal{X}$ of all mass functions on $\mathcal{X}$:

$$\Sigma_\mathcal{X} := \left\{ p \in \mathbb{R}_+^{\mathcal{X}} : \sum_{x \in \mathcal{X}} p(x) = 1 \right\}.$$
Basic concept

Consider a directed tree, with a variable $X_i$ attached to each node $i$.

Bayesian networks
The special case of a tree
Local uncertainty model associated with each node $i$

- the variable $X_i$ may assume a value in the finite set $\mathcal{X}_i$;
- for each possible value $x_{m(i)} \in \mathcal{X}_{m(i)}$ of the mother variable $X_{m(i)}$, we have a conditional mass function $p_i(\cdot| x_{m(i)})$ in $\Sigma_{\mathcal{X}_i}$:
  
  \[ p_i(z_i|x_{m(i)}) = \text{probability that } X_i = z_i, \text{ given that } X_{m(i)} = x_{m(i)} \]
- in summary: local model $p_i(\cdot| X_{m(i)})$
How should we interpret the graphical structure?

Conditional on the mother variable $X_{m(i)}$, a variable $X_i$ (and its descendants) is independent from its non-parent non-descendants.
Bayesian networks

Consequences of these independence assumptions

Consequences

1. It is a straightforward matter to infer the joint mass function from the local ones.
Bayesian networks

Consequences of these independence assumptions

Consequences

1. It is a straightforward matter to infer the joint mass function from the local ones.

2. There are very efficient message passing algorithms (Pearl) for treating the Bayesian network as an expert system: finding an updated belief model (mass function) for a queried variable after instantiating a number of other variables.
Imprecise probability models

Credal sets

I will give only a short overview, for more details see [3].

**Definition**

A credal set $\mathcal{M}$ is a **convex closed** subset of $\Sigma_X$. 
Imprecise probability models

Lower and upper expectations

\[ \Sigma_{\mathcal{X}} \]
\[ \bar{E}(I_{\{c\}}) = \frac{4}{7} \]
\[ E(I_{\{c\}}) = \frac{1}{4} \]

**Equivalent model**

Consider the set \( \mathcal{L}(\mathcal{X}) = \mathbb{R}^\mathcal{X} \) of all real-valued maps on \( \mathcal{X} \). We define two real functionals on \( \mathcal{L}(\mathcal{X}) \): for all \( f : \mathcal{X} \to \mathbb{R} \)

\[
\begin{align*}
E_M(f) &= \min \left\{ E_p(f) : p \in M \right\} \text{ lower expectation} \\
\bar{E}_M(f) &= \max \left\{ E_p(f) : p \in M \right\} \text{ upper expectation}
\end{align*}
\]

Observe that [conjugacy]:

\[ \bar{E}_M(f) = -E_M(-f). \]
Imprecise probability models

Basic properties of lower expectations

Definition

We call a real functional $E$ on $\mathcal{L}(\mathcal{X})$ an lower expectation if it satisfies the following properties:

For all $f$ and $g$ in $\mathcal{L}(\mathcal{X})$ and all real $\lambda \geq 0$:

1. $E(f) \geq \min f$ [boundedness];
2. $E(f + g) \geq E(f) + E(g)$ [super-additivity];
3. $E(\lambda f) = \lambda E(f)$ [non-negative homogeneity].

Theorem (Other properties)

Let $E$ be a lower expectation, with conjugate upper expectation $\overline{E}$. Then for all real numbers $\mu$ and all $f$ and $g$ in $\mathcal{L}(\mathcal{X})$:

1. $E(f) \leq \overline{E}(f)$;
2. $E(f) + E(g) \leq E(f + g) \leq E(f) + \overline{E}(g) \leq \overline{E}(f + g) \leq \overline{E}(f) + \overline{E}(g)$;
3. $E(f + \mu) = E(f) + \mu$;
4. $\overline{E}(|f|) \geq |E(f)|$ and $\overline{E}(|f|) \geq |\overline{E}(f)|$. 
**Theorem (Lower Envelope Theorem)**

A real functional $E$ is a lower expectation if and only if it is the lower envelope of some credal set $\mathcal{M}$.

**Proof.**

Use $\mathcal{M} = \{ p \in \Sigma_{\mathcal{X}} : (\forall f \in \mathcal{L}(\mathcal{X}))(E_p(f) \geq E(f)) \}$. 

---

**Imprecise probability models**

Lower Envelope Theorem
Imprecise probability models

Conditioning

Suppose we have two variables $X$ and $Y$, and a joint credal set

$$\mathcal{M}_{X,Y} \subseteq \Sigma_{X \times Y}$$

whose elements are joint mass functions $p(x,y)$, or equivalently, a joint lower expectation

$$\mathbf{E} : \mathcal{L}(X \times Y) \to \mathbb{R}$$

Conditioning

Suppose we observe $Y = y$, how do we update these imprecise probability models?

$$\mathcal{M}_{X,Y} \rightarrow \mathcal{M}_{X|y} \text{ and } \mathbf{E} \rightarrow \mathbf{E}(\cdot|y)$$
Imprecise probability models

Conditioning

The probability way

Apply Bayes’ Rule to each joint mass function in the credal set $\mathcal{M}$:

$$
\mathcal{M}_{X|y} = \{p(\cdot | y) : p \in \mathcal{M}_{X,Y}\} = \left\{ \frac{p(\cdot, y)}{p(y)} : p \in \mathcal{M}_{X,Y} \text{ and } p(y) > 0 \right\}
$$

$\mathcal{M}_{X|y}$ is a new credal set.

The expectation way

Apply the Generalised Bayes Rule to the joint lower expectation $E$:

$$
E(f|y) = \max \{ \mu \in \mathbb{R} : E(I_{\{y\}}[f - \mu]) \geq 0 \}, \quad f \in \mathcal{L}(\mathcal{X})
$$

Observe that $E(f|Y)$ is a real map on $\mathcal{Y}$, whose value in $y$ is $E(f|y)$.
Imprecise probability models
Marginal extension

How to go the other way around? Suppose we have:

- a marginal model $E_Y$ on $L(Y)$
- a conditional model $E(\cdot|Y)$ on $L(X)$

How do we find a joint model $E$ on $L(X \times Y)$ with marginal $E_Y$ such that conditioning produces $E(\cdot|Y)$?

No unique solution!

Law of Iterated Expectation
The smallest (most conservative) solution is given by:

$$E(g) = E_Y(E(g(\cdot,Y)|Y)),$$

$g \in L(X \times Y)$.
Imprecise probability models
Marginal extension

How to go the other way around? Suppose we have:
- a marginal model $E_Y$ on $\mathcal{L}(Y)$
- a conditional model $E(\cdot|Y)$ on $\mathcal{L}(X)$

How do we find a joint model $E$ on $\mathcal{L}(X \times Y)$ with marginal $E_Y$ such that conditioning produces $E(\cdot|Y)$?

No unique solution!
Imprecise probability models
Marginal extension

How to go the other way around? Suppose we have:

- a marginal model $\mathbb{E}_Y$ on $\mathcal{L}(Y)$
- a conditional model $\mathbb{E}(\cdot|Y)$ on $\mathcal{L}(X)$

How do we find a joint model $\mathbb{E}$ on $\mathcal{L}(X \times Y)$ with marginal $\mathbb{E}_Y$ such that conditioning produces $\mathbb{E}(\cdot|Y)$?

No unique solution!

Law of Iterated Expectation

The smallest (most conservative) solution is given by:

$$\mathbb{E}(g) = \mathbb{E}_Y(\mathbb{E}(g(\cdot,Y)|Y)), \quad g \in \mathcal{L}(X \times Y)$$
Types of independence
Three possible definitions

Epistemic irrelevance

$X_2$ is epistemically irrelevant to $X_1$, conditional on $X_3$:

$$E(f(X_1)|X_2, X_3) = E(f(X_1)|X_3)$$

Epistemic independence

$X_1$ and $X_2$ are epistemically independent, conditional on $X_3$:

$$E(f(X_1)|X_2, X_3) = E(f(X_1)|X_3) \quad \text{and} \quad E(g(X_2)|X_1, X_3) = E(g(X_2)|X_3)$$

Strong independence

Model $E(h(X_1, X_2)|X_3)$ is a lower envelope of precise independent models.
Local uncertainty model associated with each node $i$

- the variable $X_i$ may assume a value in the finite set $\mathcal{X}_i$;
- for each possible value $x_{m(i)} \in \mathcal{X}_{m(i)}$ of the mother variable $X_{m(i)}$, we have a conditional lower expectation $E_i(\cdot|x_{m(i)}) : \mathcal{L}(\mathcal{X}_i) \rightarrow \mathbb{R}$:

$$E_i(f|x_{m(i)}) = \text{lower expectation of } f(X_i), \text{ given that } X_{m(i)} = x_{m(i)}$$

- local model $E_i(\cdot|X_{m(i)})$ is a conditional lower expectation operator
- equivalent local model: credal sets $\mathcal{M}_{X_i|x_{m(i)}} \subseteq \Sigma \mathcal{X}_i$ for all $x_{m(i)} \in \mathcal{X}_{m(i)}$
Credal trees under strong independence
Interpretation of the graphical structure

How should we interpret the graphical structure?

Conditional on the mother variable $X_{m(i)}$, a variable $X_i$ (and its descendants) is strongly independent from its non-parent non-descendants.
This is the only type of credal tree (net) that has received any serious attention so far.

A credal net under strong independence is equivalent to a collection of Bayesian nets, and inferences are done by doing the inferences on these Bayesian nets, and then taking lower envelopes.

The equivalent collection of Bayesian nets is found by considering combinations of extreme points of the local credal sets: size is essentially exponential in the number of nodes.
Credal trees under strong independence

Observations

- This is the only type of credal tree (net) that has received any serious attention so far.

- A credal net under strong independence is equivalent to a collection of Bayesian nets, and inferences are done by doing the inferences on these Bayesian nets, and then taking lower envelopes.

- The equivalent collection of Bayesian nets is found by considering combinations of extreme points of the local credal sets: size is essentially exponential in the number of nodes.

- While doing inferences in Bayesian trees is very efficient, the corresponding inferences in credal trees are NP-hard.

- Exact and approximating algorithms have been devised by Marco Zaffalon’s group, and by Fabio Cozman’s team, recently also by Cassio de Campos.
Definition

The graphical structure is interpreted as follows:

Conditional on the mother variable, the non-parent non-descendants of each node variable are epistemically irrelevant to it and its descendants.
Credal trees under epistemic irrelevance

Example

\[ X_1 \rightarrow X_2 \rightarrow X_4 \rightarrow X_3 \]

- \( X_1 \) is epistemically irrelevant to \( X_3 \), conditional on \( X_2 \)
- \( X_3 \) need not be epistemically irrelevant to \( X_1 \), conditional on \( X_2 \).

Conclusion

\( X_1 \) and \( X_3 \) need not be epistemically, and certainly not strongly independent, conditional on \( X_2 \).
Credal trees under epistemic irrelevance

Example

\[ X_1 \rightarrow X_2 \rightarrow X_4 \]
\[ X_3 \rightarrow X_2 \rightarrow X_4 \]

- \( X_3 \) is epistemically irrelevant to \( X_4 \), conditional on \( X_2 \)
- \( X_4 \) is epistemically irrelevant to \( X_3 \), conditional on \( X_2 \).

Conclusion

\( X_3 \) and \( X_4 \) are epistemically, but not necessarily strongly, independent, conditional on \( X_2 \).
Credal trees under epistemic irrelevance

Some separation properties

Conclusion

For a variable $T$ to be separated from $I_1$ by a variable $I_2$, arrows should point from $I_2$ to $T$. 
Credal networks under epistemic irrelevance

As an expert system

When the credal network is a (Markov) tree we can find the joint model from the local models recursively, from leaves to root.

Exact message passing algorithm

- credal tree treated as an expert system
- linear complexity in the number of nodes

Python code

- written by Filip Hermans
- testing and connection with strong independence in cooperation with Marco Zaffalon and Alessandro Antonucci

Current (toy) applications in HMMs

- character recognition [2]
- air traffic trajectory tracking and identification [1]
Credal networks under epistemic irrelevance

HMMs: character recognition for Dante’s Divina Commedia

Original text: \[ \ldots \ V \ I \ T \ A \]

OCR output: \[ \ldots \ V \ I \ T \ O \]
Credal networks under epistemic irrelevance
HMMs: character recognition for Dante’s Divina Commedia

<table>
<thead>
<tr>
<th></th>
<th>Precise</th>
<th>Imprecise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>93.96% (7275/7743)</td>
<td>64.97% (243/374)</td>
</tr>
<tr>
<td>Determinacy</td>
<td>95.17% (7369/7743)</td>
<td></td>
</tr>
<tr>
<td>Set-accuracy</td>
<td>93.58% (350/374)</td>
<td></td>
</tr>
<tr>
<td>Single accuracy</td>
<td>95.43% (7032/7369)</td>
<td></td>
</tr>
<tr>
<td>Indeterminate output size</td>
<td>2.97</td>
<td>over 21</td>
</tr>
</tbody>
</table>

Table: Precise vs. imprecise HMMs. Test results obtained by twofold cross-validation on the first two chants of Dante’s *Divina Commedia* and $n = 2$. Quantification is achieved by IDM with $s = 2$ and modified Perks’ prior. The single-character output by the precise model is then guaranteed to be included in the set of characters the imprecise HMM identifies.
An example
A particular Markov tree

We are looking for:

\[ E(f(X_4) | x_1, x_2, x_5, x_7) \]
An example
A particular Markov tree

This is the unique $\mu$ such that:

$$E \left( \left[ f(\mathbf{X}) - \mu \right] \mathbf{I}_{\{x_1\}} \mathbf{I}_{\{x_2\}} \mathbf{I}_{\{x_5\}} \mathbf{I}_{\{x_7\}} \right) = 0$$
An example
A particular Markov tree

This is the unique $\mu$ such that:

$$E \left( [f(X_4) - \mu] I_{\{x_5\}} I_{\{x_7\}} \mid x_1 \right) = 0$$
An example
A particular Markov tree

This is the unique $\mu$ such that:

$$E_3 \left( E_4 \left( [f(X_4) - \mu] \right) \right) = 0$$
An example
A particular Markov tree

This is the unique \( \mu \) such that:

\[
E_3 \left( E_4 \left( [f(X_4) - \mu] E_5 (\{x_5\} | X_4) E_7 (\{x_7\} | X_4) | X_3 \right) \right) | x_1 = 0
\]

passed to the backbone

\[
E_3 \left( E_4 \left( [f(X_4) - \mu] E_5 (\{x_5\} | X_4) E_7 (\{x_7\} | X_4) | X_3 \right) \right) \]

passed along the backbone
Conclusion

On the positive side

- very efficient, essentially linear in number of nodes
- not much more conservative than strong independence:
  - the same for forward inference
  - dilation for backward inference

On the negative side

- the algorithm works only for trees, not for polytrees or more general acyclic directed nets
- the algorithm works only for one function $f$ at a time, it doesn’t provide the entire credal set
Alessandro Antonucci, Alessio Benavoli, Marco Zaffalon, Gert de Cooman, and Filip Hermans.
Multiple model tracking by imprecise Markov trees.
Accepted for publication.

Gert de Cooman, Filip Hermans, Alessandro Antonucci, and Marco Zaffalon.
Accepted for publication.

P. Walley.
*Statistical Reasoning with Imprecise Probabilities.*