Axiomatization of Maximum Entropy via Inductive Reasoning

Alexis Akira Toda

Department of Economics, Yale University

July 13, 2011

Questions

1. Are maximum entropy (MaxEnt) and Bayesian inference compatible with each other?

Questions

1. Are maximum entropy (MaxEnt) and Bayesian inference compatible with each other?
2. If so, which is more fundamental?

Answer to Q1

- YES. According to Jaynes (2003), MaxEnt sets up prior, subsequent inference by Bayes, hence no contradiction.

Answer to Q1

- YES. According to Jaynes (2003), MaxEnt sets up prior, subsequent inference by Bayes, hence no contradiction.
- YES. Bayes's rule implies MaxEnt (more precisely, minimum Kullback-Leibler information principle, KLIP hereafter) asymptotically (Van Campenhout & Cover 1981, Csiszár 1984).

Answer to Q1

- YES. According to Jaynes (2003), MaxEnt sets up prior, subsequent inference by Bayes, hence no contradiction.
- YES. Bayes's rule implies MaxEnt (more precisely, minimum Kullback-Leibler information principle, KLIP hereafter) asymptotically (Van Campenhout & Cover 1981, Csiszár 1984).
- YES. Both MaxEnt and Bayesian inference special cases of KLIP (Caticha & Giffin 2006).

Answer to Q1

- YES. According to Jaynes (2003), MaxEnt sets up prior, subsequent inference by Bayes, hence no contradiction.
- YES. Bayes's rule implies MaxEnt (more precisely, minimum Kullback-Leibler information principle, KLIP hereafter) asymptotically (Van Campenhout & Cover 1981, Csiszár 1984).
- YES. Both MaxEnt and Bayesian inference special cases of KLIP (Caticha & Giffin 2006).
- In the past there were a few papers mentioning the tension between MaxEnt and Bayes, but they were solving different problems.

Answer to Q2?

- Given the result of Caticha & Giffin (2006), KLIP seems most fundamental principle of inference, but is it really so?

Answer to Q2?

- Given the result of Caticha & Giffin (2006), KLIP seems most fundamental principle of inference, but is it really so?
- All axiomatizations of entropy or MaxEnt (*e.g.*, Shannon (1948), Shore & Johnson (1980), Caticha & Giffin (2006)) depend on Bayes's rule or its special case, "independence".
- Example from Shannon (1948):

$$H(p_1, p_2, p_3) = H(p_1, p_2+p_3)+(p_2+p_3)H\left(\frac{p_2}{p_2+p_3}, \frac{p_3}{p_2+p_3}\right).$$

Answer to Q2?

One view

Bayes more fundamental because Bayes's rule axiomatized by Cox (1946) and Jaynes (2003) in a very compelling way. Then KLIP follows by van Campenhout & Cover (1981).

Another view

KLIP more fundamental because it can be applied in situations with general constraints (not just moment constraints, also data; see Caticha & Giffin (2006)). Also, Bayes cannot interpret Lagrange multipliers, whereas in KLIP they are "shadow prices".

Answer to Q2?

One view

Bayes more fundamental because Bayes's rule axiomatized by Cox (1946) and Jaynes (2003) in a very compelling way. Then KLIP follows by van Campenhout & Cover (1981).

Another view

KLIP more fundamental because it can be applied in situations with general constraints (not just moment constraints, also data; see Caticha & Giffin (2006)). Also, Bayes cannot interpret Lagrange multipliers, whereas in KLIP they are "shadow prices".

I am still ambivalent between the two views, but I axiomatize KLIP without using Bayes or independence to avoid tautology.

Entropy, K-L information

- $p = \{ p_i \}$: prior, $q = \{ q_i \}$: posterior.
- Entropy (Shannon 1948):

$$H(p) = -\sum p_i \log p_i .$$

- Kullback-Leibler information (Kullback & Leibler 1951):

$$H(q; p) = \sum q_i \log \frac{q_i}{p_i} .$$

- Entropy is K-L information for uniform prior (with minus sign and additive constant).

MEP Maximum Entropy Principle (Jaynes 1957)

KLIP Minimum Kullback-Leibler Information Principle (Kullback 1959)

Jaynes's Axioms of Inductive Reasoning

Jaynes "Probability Theory: Logic of Science" (2003)

1. Degrees of plausibility are represented by real numbers.
2. Qualitative correspondence with common sense.
   (More on this next slide.)
3. Consistency.
   1. If a conclusion can be reasoned out in 2 ways, the results should be the same.
   2. The decision maker (DM) takes into account all of relevant evidence. DM is completely nonideological.
   3. DM always represents equivalent states of knowledge by equivalent plausibility assignments.
      (Laplace's principle of indifference.)

Qualitative correspondence with common sense

- Reverse monotonicity by negation:

$$p(A|C') > p(A|C) \Longrightarrow p(\neg A|C') < p(\neg A|C).$$

- Monotonicity preserved by logical conjunction:

$$\left.\begin{array}{r} p(A|C') > p(A|C) \\ p(B|A \wedge C') = p(B|A \wedge C) \end{array}\right\} \Longrightarrow p(A \wedge B|C') \geq p(A \wedge B|C).$$

Axioms of Information Gain

$p(A) \in \mathbb{R}_+$: plausibility of proposition A.

1. Numerical representation: the information gain I is a function of prior plausibility p and posterior plausibility q.
2. Continuity and monotonicity: the information gain is a continuous, increasing function in posterior plausibility.
3. Path independence: the total information gain of updating the prior plausibility p to the posterior q is independent of the path it is updated. If two paths $p \to r \to q$ and $p \to r' \to q$, then $I(p, r) + I(r, q) = I(p, r') + I(r', q)$.
4. Independence from the choice of unit: $I(tp, tq) = I(p, q)$ for $t > 0$.
5. Zero information gain for not updating: $\forall p,\ I(p, p) = 0$.

Functional Form of Information Gain

Proposition

Under axioms 1–5, information gain has the form

$$I(p, q) = k \log \frac{q}{p},$$

where $k > 0$ is an arbitrary constant. (Set $k = 1$.)

- This result, information gain $= \log \frac{\text{posterior}}{\text{prior}}$, *defined* by Goldman (1953), while I derived it.
- Ex post average information gain,

$$\sum qI(p, q) = \sum q \log \frac{q}{p},$$

is exactly K-L information.

New Axioms of Inductive Reasoning

1. Degrees of plausibility are represented by probabilities. (Finitely additive measure: don't use any probabilistic concepts.)
2. The decision maker (DM) takes into account all of relevant evidence. DM is completely nonideological.
3. Aristotelian logic: DM assigns zero plausibility to propositions that contradict his knowledge.
4. Maximum conservatism: given prior plausibilities, DM updates the plausibilities by minimizing the average information gain of the posterior plausibilities subject to known information.

Implication of New Axioms

Theorem

Under New Axioms,

1. *DM employs minimum K-L information principle (KLIP),*
2. *minimum K-L information principle implies*
   - *Jaynes's axioms, in particular the Bayes rule,*
   - *maximum likelihood,*

   and it is consistent
   (in the sense that it leads to no contradiction).

"KLIP $\Rightarrow$ Bayes"

- I: background information, $\{ A_i \}, B$: propositions.
  $\{ A_i \}$ mutually exclusive and exhaustive.
  $p(A_i \cap A_j | I)$, $p(A_i \cap B | I), p(A_i \cap B^c | I)$, etc. well defined.

"KLIP ⇒ Bayes"

- I: background information, $\{A_i\}, B$: propositions.
  $\{A_i\}$ mutually exclusive and exhaustive.
  $p(A_i \cap A_j|I)$, $p(A_i \cap B|I), p(A_i \cap B^c|I)$, etc. well defined.
- Given B, DM updates by solving

$$\min_q \sum q \log \frac{q}{p} \quad \text{s.t.} \quad (\forall i)\ q(A_i \cap B^c|B \cap I) = 0$$

$$\sum_{i=1}^{n} (q(A_i \cap B|B \cap I) + q(A_i \cap B^c|B \cap I)) = 1.$$

"KLIP ⇒ Bayes"

- I: background information, $\{A_i\}, B$: propositions.
  $\{A_i\}$ mutually exclusive and exhaustive.
  $p(A_i \cap A_j|I)$, $p(A_i \cap B|I), p(A_i \cap B^c|I)$, etc. well defined.
- Given B, DM updates by solving

$$\min_q \sum q \log \frac{q}{p} \quad \text{s.t.} \quad (\forall i)\ q(A_i \cap B^c|B \cap I) = 0$$

$$\sum_{i=1}^{n} (q(A_i \cap B|B \cap I) + q(A_i \cap B^c|B \cap I)) = 1.$$

- Using Lagrange multiplier technique, $q_i = q(A_i|B \cap I)$ satisfies

$$q(A_i|B \cap I) = q(A_i \cap B|B \cap I) + q(A_i \cap B^c|B \cap I)$$

$$= q(A_i \cap B|B \cap I) = \frac{p(A_i \cap B|I)}{\sum_{i=1}^{n} p(A_i \cap B|I)} = \frac{p(A_i \cap B|I)}{p(B|I)}.$$

"KLIP $\Rightarrow$ Maximum Likelihood"

- $\{ x_n \}$: data, $f(x)$: true density (unknown), $f(x;\theta)$: model.

"KLIP $\Rightarrow$ Maximum Likelihood"

- $\{ x_n \}$: data, $f(x)$: true density (unknown), $f(x;\theta)$: model.
- By Law of Large Numbers, K-L information is

$$
\begin{aligned}
H(f; f_\theta) &= \int f(x) \log \frac{f(x)}{f(x;\theta)} \mathrm{d}x \\
&= \int f \log f - \int f \log f_\theta = \int f \log f - \mathrm{E}_f[\log f(X;\theta)] \\
&\approx \int f \log f - \frac{1}{N} \sum_{n=1}^{N} \log f(x_n;\theta),
\end{aligned}
$$

so DM should maximize log-likelihood.

Conclusion

- Both Minimum Kullback-Leibler information principle (KLIP) and Bayes axiomatized by inductive reasoning, independent from probabilistic concepts.
- Can we replace the "independence axiom" of Caticha & Giffin (2006) by a non-probabilistic axiom similar to my approach?

My work in economics

- Moment conditions arise in economic contexts (demand = supply, Euler equation $u'(c_t) = \beta\, \mathrm{E}[u'(c_{t+1})]$).
- I apply KLIP to infer the distribution of economic variables.
- Interpret Lagrange multiplier as price.