Unreliable Probabilities and Statistical Learning

Marco Cattaneo
Department of Statistics, LMU Munich

IPSP 2014, Munich, Germany
28 June 2014

imprecise probabilities

- the uncertain beliefs of a Bayesian agent b about the state of the world $\omega \in \Omega$ are described by a (finitely additive) probability measure P_b, which is updated to $P_b(\cdot \mid A)$ by "Bayes' rule" when an event $A \subseteq \Omega$ is observed

imprecise probabilities

- the uncertain beliefs of a Bayesian agent b about the state of the world $\omega \in \Omega$ are described by a (finitely additive) probability measure P_b, which is updated to $P_b(\cdot \mid A)$ by "Bayes' rule" when an event $A \subseteq \Omega$ is observed

- an imprecise probability model $\mathcal{P} = \{P_b : b \in \mathcal{B}\}$ can be seen as a group $\mathcal{B}$ of Bayesian agents deciding by unanimity, but otherwise **not interacting**

imprecise probabilities

- the uncertain beliefs of a Bayesian agent b about the state of the world $\omega \in \Omega$ are described by a (finitely additive) probability measure P_b, which is updated to $P_b(\cdot \mid A)$ by "Bayes' rule" when an event $A \subseteq \Omega$ is observed

- an imprecise probability model $\mathcal{P} = \{P_b : b \in \mathcal{B}\}$ can be seen as a group $\mathcal{B}$ of Bayesian agents deciding by unanimity, but otherwise **not interacting**

- in particular, $\mathcal{P}$ is updated to $\{P_b(\cdot \mid A) : b \in \mathcal{B}\}$ by "generalized Bayes' rule" when an event A is observed

unreliable probabilities

- Gärdenfors and Sahlin (1982) proposed a hierarchical model consisting of $\mathcal{P}$ (first level) and a measure ρ of **reliability/credibility** of the Bayesian agents $b \in \mathcal{B}$ (second level)

unreliable probabilities

- Gärdenfors and Sahlin (1982) proposed a hierarchical model consisting of $\mathcal{P}$ (first level) and a measure ρ of **reliability/credibility** of the Bayesian agents $b \in \mathcal{B}$ (second level)
- the hierarchical model generalizes the imprecise probability model (corresponding to the case in which all Bayesian agents are equally reliable/credible), but the second-order "measure" ρ does not have a clear interpretation or mathematical form

unreliable probabilities

- Gärdenfors and Sahlin (1982) proposed a hierarchical model consisting of $\mathcal{P}$ (first level) and a measure ρ of **reliability/credibility** of the Bayesian agents $b \in \mathcal{B}$ (second level)

- the hierarchical model generalizes the imprecise probability model (corresponding to the case in which all Bayesian agents are equally reliable/credible), but the second-order "measure" ρ does not have a clear interpretation or mathematical form

- examples of similar models:

unreliable probabilities

- Gärdenfors and Sahlin (1982) proposed a hierarchical model consisting of $\mathcal{P}$ (first level) and a measure ρ of **reliability/credibility** of the Bayesian agents $b \in \mathcal{B}$ (second level)

- the hierarchical model generalizes the imprecise probability model (corresponding to the case in which all Bayesian agents are equally reliable/credible), but the second-order "measure" ρ does not have a clear interpretation or mathematical form

- examples of similar models:
  - ρ is a **possibility measure** with no clear interpretation (Zadeh, 1984; Buckley, 2003)

unreliable probabilities

- Gärdenfors and Sahlin (1982) proposed a hierarchical model consisting of $\mathcal{P}$ (first level) and a measure ρ of **reliability/credibility** of the Bayesian agents $b \in \mathcal{B}$ (second level)

- the hierarchical model generalizes the imprecise probability model (corresponding to the case in which all Bayesian agents are equally reliable/credible), but the second-order "measure" ρ does not have a clear interpretation or mathematical form

- examples of similar models:
  - ρ is a **possibility measure** with no clear interpretation (Zadeh, 1984; Buckley, 2003)
  - ρ is a probability measure (Good, 1965; Sahlin, 1983)

unreliable probabilities

- Gärdenfors and Sahlin (1982) proposed a hierarchical model consisting of $\mathcal{P}$ (first level) and a measure ρ of **reliability/credibility** of the Bayesian agents $b \in \mathcal{B}$ (second level)

- the hierarchical model generalizes the imprecise probability model (corresponding to the case in which all Bayesian agents are equally reliable/credible), but the second-order "measure" ρ does not have a clear interpretation or mathematical form

- examples of similar models:
  - ρ is a **possibility measure** with no clear interpretation (Zadeh, 1984; Buckley, 2003)
  - ρ is a probability measure (Good, 1965; Sahlin, 1983)
  - ρ is a **possibility measure** with an upper probability interpretation (Walley, 1997; de Cooman, 2005)

statistical learning

- when an event A is observed, the "generalized Bayes' rule" **discards the information** in A for discrimination between $b, b' \in \mathcal{B}$ (Kullback and Leibler, 1951), or weight of evidence in favor of b against b' (Good, 1950):

$$\log \frac{P_b(A)}{P_{b'}(A)}$$

statistical learning

- when an event A is observed, the "generalized Bayes' rule" **discards the information** in A for discrimination between $b, b' \in \mathcal{B}$ (Kullback and Leibler, 1951), or weight of evidence in favor of b against b' (Good, 1950):

$$\log \frac{P_b(A)}{P_{b'}(A)}$$

- this information is summarized by the (second-order) likelihood function $\lambda_A : b \mapsto P_b(A)$, which would be used to update a second-order probability measure ρ (precise or imprecise)

statistical learning

- when an event A is observed, the "generalized Bayes' rule" **discards the information** in A for discrimination between $b, b' \in \mathcal{B}$ (Kullback and Leibler, 1951), or weight of evidence in favor of b against b' (Good, 1950):

$$\log \frac{P_b(A)}{P_{b'}(A)}$$

- this information is summarized by the (second-order) likelihood function $\lambda_A : b \mapsto P_b(A)$, which would be used to update a second-order probability measure ρ (precise or imprecise)

- the likelihood function λ_A describes the (relative) ability of the Bayesian agents to predict the event A

example: coin tossing

- a particular coin is known to be either fair or loaded with a $\frac{3}{4}$ probability for one of the two sides

example: coin tossing

- a particular coin is known to be either fair or loaded with a $\frac{3}{4}$ probability for one of the two sides
- the Bayesian agent b believes that the coin is either fair or loaded toward heads (with the same prior probability $\frac{1}{2}$ for these two possibilities), while the Bayesian agent b' believes that the coin is either fair or loaded toward tails (with the same prior probability $\frac{1}{2}$ for these two possibilities):

$$P_b(\text{heads in the next toss}) = 0.625$$
$$P_{b'}(\text{heads in the next toss}) = 0.375$$

example: coin tossing

- a particular coin is known to be either fair or loaded with a $\frac{3}{4}$ probability for one of the two sides
- the Bayesian agent b believes that the coin is either fair or loaded toward heads (with the same prior probability $\frac{1}{2}$ for these two possibilities), while the Bayesian agent b' believes that the coin is either fair or loaded toward tails (with the same prior probability $\frac{1}{2}$ for these two possibilities):

$$P_b(\text{heads in the next toss}) = 0.625$$
$$P_{b'}(\text{heads in the next toss}) = 0.375$$

- the event $A = \{77 \text{ heads in the first } 100 \text{ tosses}\}$ is observed:

$$P_b(\text{heads in the next toss} \,|\, A) \approx 0.745$$
$$P_{b'}(\text{heads in the next toss} \,|\, A) \approx 0.500$$

example: coin tossing

- a particular coin is known to be either fair or loaded with a $\frac{3}{4}$ probability for one of the two sides
- the Bayesian agent b believes that the coin is either fair or loaded toward heads (with the same prior probability $\frac{1}{2}$ for these two possibilities), while the Bayesian agent b' believes that the coin is either fair or loaded toward tails (with the same prior probability $\frac{1}{2}$ for these two possibilities):

$$P_b(\text{heads in the next toss}) = 0.625$$
$$P_{b'}(\text{heads in the next toss}) = 0.375$$

- the event $A = \{77 \text{ heads in the first } 100 \text{ tosses}\}$ is observed:

$$P_b(\text{heads in the next toss} \,|\, A) \approx 0.745$$
$$P_{b'}(\text{heads in the next toss} \,|\, A) \approx 0.500$$

- weight of evidence in favor of b against b':

$$\log \frac{P_b(A)}{P_{b'}(A)} = \log \frac{\lambda_A(b)}{\lambda_A(b')} \approx \log(4.32 \times 10^6) \approx 66.4\,\mathrm{db}$$

hierarchical model

- the relative reliability/credibility of the Bayesian agents $b \in \mathcal{B}$ can be interpreted as the relative quality of their past forecasts, which is described by the likelihood function λ_A (where A represents all past observations, real or imagined)

hierarchical model

- the relative reliability/credibility of the Bayesian agents $b \in \mathcal{B}$ can be interpreted as the relative quality of their past forecasts, which is described by the likelihood function λ_A (where A represents all past observations, real or imagined)

- the second-order measure ρ of (relative) reliability/credibility can thus be identified with the likelihood function λ_A (Cattaneo, 2008), or with its normalized extension to subsets $\mathcal{S} \subseteq \mathcal{B}$: the likelihood ratio

$$\Lambda_A : \mathcal{S} \mapsto \frac{\sup_{b \in \mathcal{S}} \lambda_A(b)}{\sup_{b' \in \mathcal{B}} \lambda_A(b')}$$

hierarchical model

- the relative reliability/credibility of the Bayesian agents $b \in \mathcal{B}$ can be interpreted as the relative quality of their past forecasts, which is described by the likelihood function λ_A (where A represents all past observations, real or imagined)

- the second-order measure ρ of (relative) reliability/credibility can thus be identified with the likelihood function λ_A (Cattaneo, 2008), or with its normalized extension to subsets $\mathcal{S} \subseteq \mathcal{B}$: the likelihood ratio

$$\Lambda_A : \mathcal{S} \mapsto \frac{\sup_{b \in \mathcal{S}} \lambda_A(b)}{\sup_{b' \in \mathcal{B}} \lambda_A(b')}$$

- Λ_A is a **possibility measure**, whose updating rule (unlike the ones of similar models with second-order possibility measures) seems to fit with the informal description of Gärdenfors and Sahlin (1982): $\mathcal{P}$ is updated by "generalized Bayes' rule" and Λ_A is updated to $\Lambda_{A \cap B}$ when an event B is observed

complete ignorance

- a constant likelihood function λ_A describes the case of **no information** for discrimination among the Bayesian agents $b \in \mathcal{B}$ (very intuitive idea): in this case, the possibility measure Λ_A is the vacuous upper probability measure on $\mathcal{B}$ (complete ignorance about b implies complete ignorance about $f(b)$, for all functions f)

complete ignorance

- a constant likelihood function λ_A describes the case of **no information** for discrimination among the Bayesian agents $b \in \mathcal{B}$ (very intuitive idea): in this case, the possibility measure Λ_A is the vacuous upper probability measure on $\mathcal{B}$ (complete ignorance about b implies complete ignorance about $f(b)$, for all functions f)

- basic advantage of the hierarchical model over:

complete ignorance

- a constant likelihood function λ_A describes the case of **no information** for discrimination among the Bayesian agents $b \in \mathcal{B}$ (very intuitive idea): in this case, the possibility measure Λ_A is the vacuous upper probability measure on $\mathcal{B}$ (complete ignorance about b implies complete ignorance about $f(b)$, for all functions f)

- basic advantage of the hierarchical model over:
  - the Bayesian model: the ability to **describe** the state of complete ignorance

complete ignorance

- a constant likelihood function λ_A describes the case of **no information** for discrimination among the Bayesian agents $b \in \mathcal{B}$ (very intuitive idea): in this case, the possibility measure Λ_A is the vacuous upper probability measure on $\mathcal{B}$ (complete ignorance about b implies complete ignorance about $f(b)$, for all functions f)

- basic advantage of the hierarchical model over:
  - the Bayesian model: the ability to **describe** the state of complete ignorance
  - the imprecise probability model: the ability to **get out of** the state of complete ignorance

complete ignorance

- a constant likelihood function λ_A describes the case of **no information** for discrimination among the Bayesian agents $b \in \mathcal{B}$ (very intuitive idea): in this case, the possibility measure Λ_A is the vacuous upper probability measure on $\mathcal{B}$ (complete ignorance about b implies complete ignorance about $f(b)$, for all functions f)

- basic advantage of the hierarchical model over:
  - the Bayesian model: the ability to **describe** the state of complete ignorance
  - the imprecise probability model: the ability to **get out of** the state of complete ignorance

- for the imprecise probability model, the state of complete ignorance corresponds to a group of Bayesian agents who are absolutely certain of different things (there is no lack of information: on the contrary, there is plenty of contradictory information), while for the hierarchical model the state of complete ignorance corresponds to the lack of information for evaluating the reliability/credibility of these agents

conclusion

- some advantages of the hierarchical model over the imprecise probability model:

conclusion

- some advantages of the hierarchical model over the imprecise probability model:
  - generality (the Bayesian agents do not have to be equally reliable/credible)

conclusion

- some advantages of the hierarchical model over the imprecise probability model:
  - generality (the Bayesian agents do not have to be equally reliable/credible)
  - ability to get out of the state of **complete ignorance**

conclusion

- some advantages of the hierarchical model over the imprecise probability model:
  - generality (the Bayesian agents do not have to be equally reliable/credible)
  - ability to get out of the state of **complete ignorance**
  - connection with classical statistics (**repeated sampling** properties of likelihood methods)

conclusion

- some advantages of the hierarchical model over the imprecise probability model:
  - generality (the Bayesian agents do not have to be equally reliable/credible)
  - ability to get out of the state of **complete ignorance**
  - connection with classical statistics (**repeated sampling** properties of likelihood methods)
  - **continuity** of updating rule (Cattaneo, 2014)

conclusion

- some advantages of the hierarchical model over the imprecise probability model:
  - generality (the Bayesian agents do not have to be equally reliable/credible)
  - ability to get out of the state of **complete ignorance**
  - connection with classical statistics (**repeated sampling** properties of likelihood methods)
  - **continuity** of updating rule (Cattaneo, 2014)
  - manageability (reduction of imprecision, information fusion, ...)

conclusion

- some advantages of the hierarchical model over the imprecise probability model:
  - generality (the Bayesian agents do not have to be equally reliable/credible)
  - ability to get out of the state of **complete ignorance**
  - connection with classical statistics (**repeated sampling** properties of likelihood methods)
  - **continuity** of updating rule (Cattaneo, 2014)
  - manageability (reduction of imprecision, information fusion, . . .)
- a drawback of the hierarchical model is the lack of a justification of the updating rule in terms of coherence or avoidance of **sure loss**

conclusion

- some advantages of the hierarchical model over the imprecise probability model:
  - generality (the Bayesian agents do not have to be equally reliable/credible)
  - ability to get out of the state of **complete ignorance**
  - connection with classical statistics (**repeated sampling** properties of likelihood methods)
  - **continuity** of updating rule (Cattaneo, 2014)
  - manageability (reduction of imprecision, information fusion, . . .)
- a drawback of the hierarchical model is the lack of a justification of the updating rule in terms of coherence or avoidance of **sure loss**
- conflict between statistical learning and behaviorist interpretation of updating

references

Buckley, J. J. (2003). *Fuzzy Probabilities*. Physica-Verlag.

Cattaneo, M. (2008). Fuzzy probabilities based on the likelihood function. In *Soft Methods for Handling Variability and Imprecision*. Springer, 43–50.

Cattaneo, M. (2014). A continuous updating rule for imprecise probabilities. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems, Part 3*. Springer, 426–435.

de Cooman, G. (2005). A behavioural model for vague probability assessments. *Fuzzy Sets Syst.* 154, 305–358.

Gärdenfors, P., and Sahlin, N.-E. (1982). Unreliable probabilities, risk taking, and decision making. *Synthese* 53, 361–386.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. Charles Griffin.

Good, I. J. (1965). *The Estimation of Probabilities*. MIT Press.

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.* 22, 79–86.

Sahlin, N.-E. (1983). On second order probabilities and the notion of epistemic risk. In *Foundations of Utility and Risk Theory with Applications*. Springer, 95–104.

Walley, P. (1997). Statistical inferences based on a second-order possibility distribution. *Int. J. Gen. Syst.* 26, 337–383.

Zadeh, L. A. (1984). Fuzzy probabilities. *Inf. Process. Manage.* 20, 363–372.