A hierarchical model based on the likelihood function

Marco Cattaneo
Department of Statistics, LMU Munich
cattaneo@stat.uni-muenchen.de

May 12, 2008

my research

- PhD with Frank Hampel at ETH Zurich
  (November 2002 – March 2007):

  Statistical Decisions Based Directly on the Likelihood Function

my research

- PhD with Frank Hampel at ETH Zurich
  (November 2002 – March 2007):

  Statistical Decisions Based Directly on the Likelihood Function

- Postdoc with Thomas Augustin at LMU Munich
  (SNSF Research Fellowship, October 2007 – September 2008):

  Decision making on the basis of a probabilistic-possibilistic hierarchical description of uncertain knowledge

the likelihood function

Let $\mathcal{P}$ be a set of probability measures on a measurable space $(\Omega, \mathcal{A})$.

Each $P \in \mathcal{P}$ is interpreted as a probabilistic model of the reality under consideration. The interpretation of probability is not important: for instance the elements of $\mathcal{P}$ can be statistical models, or describe the forecasts of a group of experts.

the likelihood function

Let $\mathcal{P}$ be a set of probability measures on a measurable space $(\Omega, \mathcal{A})$.

Each $P \in \mathcal{P}$ is interpreted as a probabilistic model of the reality under consideration. The interpretation of probability is not important: for instance the elements of $\mathcal{P}$ can be statistical models, or describe the forecasts of a group of experts.

When an event $A \in \mathcal{A}$ is observed, the (normalized) **likelihood function**

$$lik : P \longmapsto \frac{P(A)}{\sup_{P \in \mathcal{P}} P(A)}$$

on $\mathcal{P}$ describes the *relative* ability of the probabilistic models in $\mathcal{P}$ to forecast the observed data.

The likelihood function can be interpreted as a measure of the *relative* plausibility of the probabilistic models in the light of the observed data alone.

running example

The **fundamental problem of practical statistics** (Pearson,1920):

> *An "event" has occurred p times out of $p + q = n$ trials, where we have no a priori knowledge of the frequency of the event in the total population of occurrences. What is the probability of its occurring r times in a further $r + s = m$ trials?*

running example

The **fundamental problem of practical statistics** (Pearson,1920):

> *An "event" has occurred p times out of $p + q = n$ trials, where we have no a priori knowledge of the frequency of the event in the total population of occurrences. What is the probability of its occurring r times in a further $r + s = m$ trials?*

classical approach:

probability measure P_θ on $(\Omega, \mathcal{A})$
such that $X_1, X_2, \ldots \overset{i.i.d.}{\sim} Ber(\theta)$
(that is, $P_\theta\{X_i = 1\} = \theta$
and $P_\theta\{X_i = 0\} = 1 - \theta$),
where $\theta \in \Theta = [0, 1]$ is the unknown
probability of the "event"

running example

The **fundamental problem of practical statistics** (Pearson,1920):

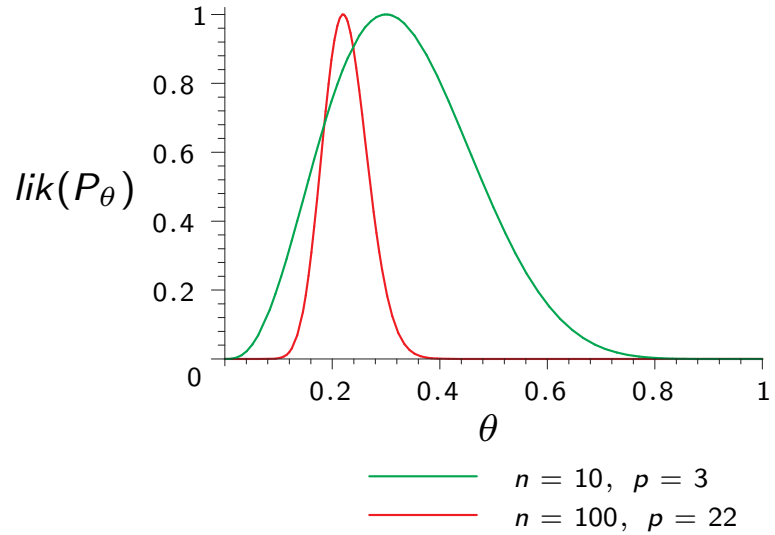
> *An "event" has occurred p times out of $p + q = n$ trials, where we have no a priori knowledge of the frequency of the event in the total population of occurrences. What is the probability of its occurring r times in a further $r + s = m$ trials?*

classical approach:

probability measure P_θ on $(\Omega, \mathcal{A})$
such that $X_1, X_2, \ldots \overset{i.i.d.}{\sim} Ber(\theta)$
(that is, $P_\theta\{X_i = 1\} = \theta$
and $P_\theta\{X_i = 0\} = 1 - \theta$),
where $\theta \in \Theta = [0, 1]$ is the unknown probability of the "event"

$\mathcal{P}_{param} = \{P_\theta : \theta \in \Theta\}$

$lik : P_\theta \mapsto \frac{n^n}{p^p\, q^q}\, \theta^p\, (1 - \theta)^q$



running example

Bayesian approach:

exchangeability and prior probability measure π on Θ,
leading to a probability measure P_π on $\Theta \times \Omega$

running example

Bayesian approach:

exchangeability and prior probability measure π on Θ,
leading to a probability measure P_π on $\Theta \times \Omega$

$\mathcal{P}_{Bayes} = \{P_\pi\}$

$lik \equiv 1$

running example

Bayesian approach:

exchangeability and prior probability measure π on Θ,
leading to a probability measure P_π on $\Theta \times \Omega$

$\mathcal{P}_{Bayes} = \{P_\pi\}$

$lik \equiv 1$

IP approach:
exchangeability and prior imprecise probability measure Π on Θ,
where Π is a set of probability measures on Θ

running example

Bayesian approach:

exchangeability and prior probability measure π on Θ,
leading to a probability measure P_π on $\Theta \times \Omega$

$\mathcal{P}_{Bayes} = \{P_\pi\}$

$lik \equiv 1$

IP approach:
exchangeability and prior imprecise probability measure Π on Θ,
where Π is a set of probability measures on Θ

$\mathcal{P}_{IP} = \{P_\pi : \pi \in \Pi\}$

$lik : P_\pi \mapsto \frac{E_\pi[\theta^p (1-\theta)^q]}{\sup_{\pi \in \Pi} E_\pi[\theta^p (1-\theta)^q]}$

running example

Bayesian approach:

exchangeability and prior probability measure π on Θ,
leading to a probability measure P_π on $\Theta \times \Omega$

$\mathcal{P}_{Bayes} = \{P_\pi\}$

$lik \equiv 1$

IP approach:
exchangeability and prior imprecise probability measure Π on Θ,
where Π is a set of probability measures on Θ

$\mathcal{P}_{IP} = \{P_\pi : \pi \in \Pi\}$

$lik : P_\pi \mapsto \frac{E_\pi[\theta^p (1-\theta)^q]}{\sup_{\pi \in \Pi} E_\pi[\theta^p (1-\theta)^q]}$

for example:

$\mathcal{P}_{ignor} = \{P_\pi : \pi \text{ prob on } \Theta\}$

running example

Bayesian approach:

exchangeability and prior probability measure π on Θ,
leading to a probability measure P_π on $\Theta \times \Omega$

$\mathcal{P}_{Bayes} = \{P_\pi\}$

$lik \equiv 1$

IP approach:
exchangeability and prior imprecise probability measure Π on Θ,
where Π is a set of probability measures on Θ

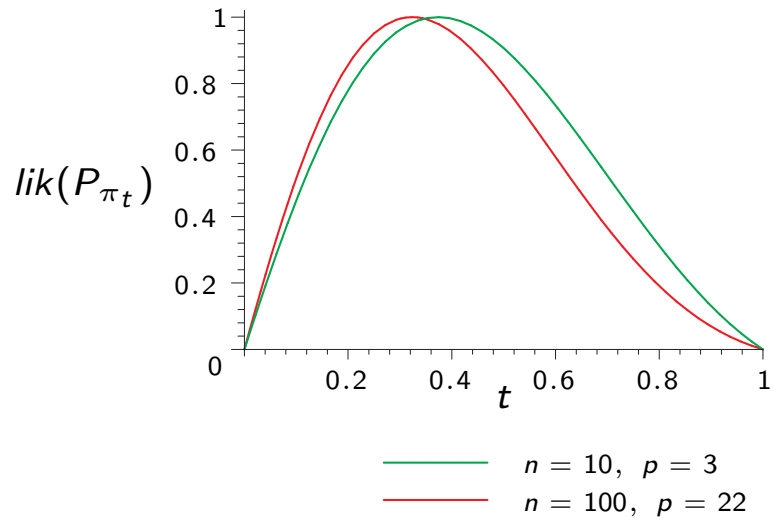
$\mathcal{P}_{IP} = \{P_\pi : \pi \in \Pi\}$

$lik : P_\pi \mapsto \frac{E_\pi[\theta^p\,(1-\theta)^q]}{\sup_{\pi \in \Pi} E_\pi[\theta^p\,(1-\theta)^q]}$

for example:

$\mathcal{P}_{ignor} = \{P_\pi : \pi \text{ prob on } \Theta\}$

$\mathcal{P}_{IDM(2)} = \{P_{\pi_t} : t \in (0,1)\}$,
where π_t is Beta$(2\,t,\, 2-2\,t)$



possibility distributions

One of the established semantics of possibility theory interprets possibility distributions as likelihood functions (see for example Hisdal, 1988).

possibility distributions

One of the established semantics of possibility theory interprets possibility distributions as likelihood functions (see for example Hisdal, 1988).

The **likelihood ratio** test discards the hypothesis that the data were generated by some probabilistic model in $\mathcal{H} \subseteq \mathcal{P}$ if

$$LR(\mathcal{H}) = \sup_{P \in \mathcal{H}} lik(P)$$

is sufficiently small. Interpreted as a set function, $LR : 2^{\mathcal{P}} \rightarrow [0, 1]$ is the possibility measure on $\mathcal{P}$ with possibility distribution lik.

possibility distributions

One of the established semantics of possibility theory interprets possibility distributions as likelihood functions (see for example Hisdal, 1988).

The **likelihood ratio** test discards the hypothesis that the data were generated by some probabilistic model in $\mathcal{H} \subseteq \mathcal{P}$ if

$$LR(\mathcal{H}) = \sup_{P \in \mathcal{H}} lik(P)$$

is sufficiently small. Interpreted as a set function, $LR : 2^{\mathcal{P}} \to [0, 1]$ is the possibility measure on $\mathcal{P}$ with possibility distribution lik.

The constant likelihood function $lik \equiv 1$ describes **complete ignorance** (in the sense of absence of information for discrimination between the probabilistic models): in this case, $LR(\mathcal{H}) = 1$ for all nonempty $\mathcal{H} \subseteq \mathcal{P}$.

the hierarchical model

The probabilistic models in $\mathcal{P}$ and the likelihood function *lik* on $\mathcal{P}$ can be interpreted as the two levels of a (probabilistic-possibilistic) **hierarchical model** of the reality under consideration.

the hierarchical model

The probabilistic models in $\mathcal{P}$ and the likelihood function *lik* on $\mathcal{P}$ can be interpreted as the two levels of a (probabilistic-possibilistic) **hierarchical model** of the reality under consideration.

The definition of likelihood function implies that when an event $A \in \mathcal{A}$ is observed, the probabilistic level $\mathcal{P}$ is updated to

$$\mathcal{P}' = \{P(\,\cdot\,|A) : P \in \mathcal{P},\, P(A) > 0\}$$

(this corresponds to the usual updating rule for IP models), and the possibilistic level *lik* is updated to

$$lik' : P' \longmapsto \frac{\sup_{P \in \mathcal{P} \,:\, P(\,\cdot\,|A) = P'} lik(P)\, P(A)}{\sup_{P \in \mathcal{P}} lik(P)\, P(A)}.$$

the hierarchical model

The probabilistic models in $\mathcal{P}$ and the likelihood function *lik* on $\mathcal{P}$ can be interpreted as the two levels of a (probabilistic-possibilistic) **hierarchical model** of the reality under consideration.

The definition of likelihood function implies that when an event $A \in \mathcal{A}$ is observed, the probabilistic level $\mathcal{P}$ is updated to

$$\mathcal{P}' = \{P(\,\cdot\,|A) : P \in \mathcal{P},\, P(A) > 0\}$$

(this corresponds to the usual updating rule for IP models), and the possibilistic level *lik* is updated to

$$lik' : P' \longmapsto \frac{\sup_{P \in \mathcal{P} \,:\, P(\,\cdot\,|A) = P'} lik(P)\, P(A)}{\sup_{P \in \mathcal{P}} lik(P)\, P(A)}.$$

When A is the first observed event, the constant likelihood function $lik \equiv 1$ describes prior ignorance, while other **prior** likelihood functions *lik* can be interpreted as (subjective) measures of the relative plausibility of the probabilistic models in $\mathcal{P}$ according to the prior information.

fuzzy expectations

The uncertain knowledge about the value $g(P)$ of a function $g : \mathcal{P} \to \mathcal{G}$ is described by the induced possibility measure $LR \circ g^{-1}$ on $\mathcal{G}$, whose possibility distribution is the **profile** likelihood function

$$lik_g : \gamma \longmapsto \sup_{P \in \mathcal{P} \,:\, g(P) = \gamma} lik(P).$$

fuzzy expectations

The uncertain knowledge about the value $g(P)$ of a function $g : \mathcal{P} \to \mathcal{G}$ is described by the induced possibility measure $LR \circ g^{-1}$ on $\mathcal{G}$, whose possibility distribution is the **profile** likelihood function

$$lik_g : \gamma \longmapsto \sup_{P \in \mathcal{P} \,:\, g(P) = \gamma} lik(P).$$

In particular, if $g : \mathcal{P} \to \mathbb{R}$ associates to each probabilistic model P the corresponding expectation $g(P) = E_P(X)$ of a random variable X, then lik_g can be interpreted as the membership function of the **fuzzy expectation** of X (**fuzzy probability** of $A \in \mathcal{A}$ when $X = I_A$).

fuzzy expectations

The uncertain knowledge about the value $g(P)$ of a function $g : \mathcal{P} \to \mathcal{G}$ is described by the induced possibility measure $LR \circ g^{-1}$ on $\mathcal{G}$, whose possibility distribution is the **profile** likelihood function

$$lik_g : \gamma \longmapsto \sup_{P \in \mathcal{P} : g(P) = \gamma} lik(P).$$

In particular, if $g : \mathcal{P} \to \mathbb{R}$ associates to each probabilistic model P the corresponding expectation $g(P) = E_P(X)$ of a random variable X, then lik_g can be interpreted as the membership function of the **fuzzy expectation** of X (**fuzzy probability** of $A \in \mathcal{A}$ when $X = I_A$).

In this case, the support $supp(lik_g) = \{x \in \mathbb{R} : lik_g(x) > 0\}$ of lik_g satisfies

$$\inf supp(lik_g) = \inf_{P \in \mathcal{P}} E_P(X) \quad \text{and} \quad \sup supp(lik_g) = \sup_{P \in \mathcal{P}} E_P(X);$$

that is, the hierarchical model generalizes the IP model by additionally considering the relative plausibility of different values in the expectations intervals (and in particular in the probability intervals).

running example

In the "fundamental problem of practical statistics", let $r = s = 1$; that is, we are interested in the (conditional) probability of $X_{n+1} + X_{n+2} = 1$.

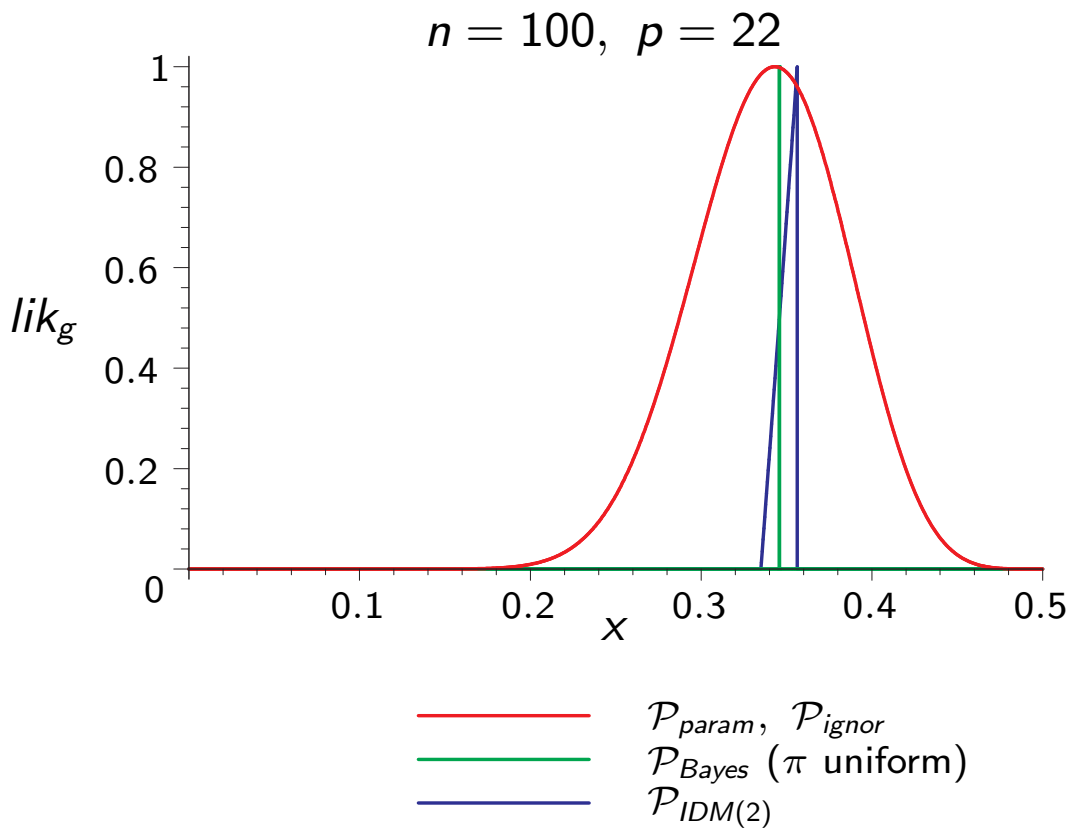
running example

In the "fundamental problem of practical statistics", let $r = s = 1$; that is, we are interested in the (conditional) probability of $X_{n+1} + X_{n+2} = 1$.

$g : P \mapsto P\{X_{n+1} + X_{n+2} = 1\}$

$\mathcal{P}_{param}$, $\mathcal{P}_{Bayes}$: the fuzzy probabilities are in agreement with the results of the classical (likelihood) and Bayesian approaches

$\mathcal{P}_{IP}$: the fuzzy probabilities contain more information than the probability intervals of the IP approach (which correspond to $supp(lik_g)$)



evaluations

It is often useful to reduce the fuzzy expectation of a random variable X to a pair of real numbers $V_*[E_P(X)]$ and $V^*[E_P(X)]$, interpreted as lower and upper evaluations.

evaluations

It is often useful to reduce the fuzzy expectation of a random variable X to a pair of real numbers $V_*[E_P(X)]$ and $V^*[E_P(X)]$, interpreted as lower and upper evaluations.

Examples of reasonable evaluations are the coherent previsions

$$\int^{\mathrm{C}} E_P(X)\,\mathrm{d}\overline{LR}(P) \quad \text{and} \quad \int^{\mathrm{C}} E_P(X)\,\mathrm{d}LR(P),$$

and the centered convex previsions (see Pelessoni and Vicig, 2003)

$$\inf_{P\in\mathcal{P}}[E_P(X) - \log lik(P)] \quad \text{and} \quad \sup_{P\in\mathcal{P}}[E_P(X) + \log lik(P)].$$

evaluations

It is often useful to reduce the fuzzy expectation of a random variable X to a pair of real numbers $V_*[E_P(X)]$ and $V^*[E_P(X)]$, interpreted as lower and upper evaluations.

Examples of reasonable evaluations are the coherent previsions

$$\int^{\mathrm{C}} E_P(X)\,\mathrm{d}\overline{LR}(P) \quad \text{and} \quad \int^{\mathrm{C}} E_P(X)\,\mathrm{d}LR(P),$$

and the centered convex previsions (see Pelessoni and Vicig, 2003)

$$\inf_{P\in\mathcal{P}}[E_P(X) - \log lik(P)] \quad \text{and} \quad \sup_{P\in\mathcal{P}}[E_P(X) + \log lik(P)].$$

When there is no information for discrimination between the probabilistic models in $\mathcal{P}$ (that is, $lik \equiv 1$), these pairs of evaluations reduce to the coherent previsions resulting from the IP model:

$$P_*[E_P(X)] = \inf_{P\in\mathcal{P}}[E_P(X)] \quad \text{and} \quad P^*[E_P(X)] = \sup_{P\in\mathcal{P}}[E_P(X)].$$

inconsistency

The usual updating rule for IP models disregards the information for discrimination between the probabilistic models in $\mathcal{P}$ (since it disregards the likelihood function *lik* on $\mathcal{P}$).

inconsistency

The usual updating rule for IP models disregards the information for discrimination between the probabilistic models in $\mathcal{P}$ (since it disregards the likelihood function *lik* on $\mathcal{P}$).

For instance, if the elements of $\mathcal{P}$ describe the opinions of a group of Bayesian experts, then the usual updating rule for IP models corresponds to update the opinion of each expert without reconsidering her/his credibility, independently of how bad her/his forecasts were when compared to the forecasts of the other experts.

inconsistency

The usual updating rule for IP models disregards the information for discrimination between the probabilistic models in $\mathcal{P}$ (since it disregards the likelihood function *lik* on $\mathcal{P}$).

For instance, if the elements of $\mathcal{P}$ describe the opinions of a group of Bayesian experts, then the usual updating rule for IP models corresponds to update the opinion of each expert without reconsidering her/his credibility, independently of how bad her/his forecasts were when compared to the forecasts of the other experts.

Since it disregards a part of the information provided by the data, the usual updating rule for IP models can lead to statistical **inconsistency** even in simple problems.

running example

A pair of lower and upper evaluations of $P\{X_{n+1} + X_{n+2} = 1\}$ are consistent if, for all $\theta \in \Theta$ that are not discarded by the prior information, they tend (in probability) to $P_\theta\{X_{n+1} + X_{n+2} = 1\} = 2\,\theta\,(1-\theta)$ as $n \longrightarrow \infty$, when $X_1, \ldots, X_n$ are distributed according to P_θ.

running example

A pair of lower and upper evaluations of $P\{X_{n+1} + X_{n+2} = 1\}$ are consistent if, for all $\theta \in \Theta$ that are not discarded by the prior information, they tend (in probability) to $P_\theta\{X_{n+1} + X_{n+2} = 1\} = 2\,\theta\,(1-\theta)$ as $n \to \infty$, when $X_1, \ldots, X_n$ are distributed according to P_θ.

hierarchical model:
all reasonable evaluations are consistent, and prior ignorance is possible

running example

A pair of lower and upper evaluations of $P\{X_{n+1} + X_{n+2} = 1\}$ are consistent if, for all $\theta \in \Theta$ that are not discarded by the prior information, they tend (in probability) to $P_\theta\{X_{n+1} + X_{n+2} = 1\} = 2\,\theta\,(1-\theta)$ as $n \to \infty$, when $X_1, \ldots, X_n$ are distributed according to P_θ.

hierarchical model:
all reasonable evaluations are consistent, and prior ignorance is possible

Bayesian model:
all posterior probabilities are consistent, but prior ignorance is impossible

running example

A pair of lower and upper evaluations of $P\{X_{n+1} + X_{n+2} = 1\}$ are consistent if, for all $\theta \in \Theta$ that are not discarded by the prior information, they tend (in probability) to $P_\theta\{X_{n+1} + X_{n+2} = 1\} = 2\,\theta\,(1-\theta)$ as $n \to \infty$, when $X_1, \ldots, X_n$ are distributed according to P_θ.

hierarchical model:
all reasonable evaluations are consistent, and prior ignorance is possible

Bayesian model:
all posterior probabilities are consistent, but prior ignorance is impossible

IP model:

$\mathcal{P}_{ignor}$: prior ignorance, but no consistency
(the probability interval is $[0, \frac{1}{2}]$ for all n)

running example

A pair of lower and upper evaluations of $P\{X_{n+1} + X_{n+2} = 1\}$ are consistent if, for all $\theta \in \Theta$ that are not discarded by the prior information, they tend (in probability) to $P_\theta\{X_{n+1} + X_{n+2} = 1\} = 2\,\theta\,(1-\theta)$ as $n \to \infty$, when $X_1, \ldots, X_n$ are distributed according to P_θ.

hierarchical model:
all reasonable evaluations are consistent, and prior ignorance is possible

Bayesian model:
all posterior probabilities are consistent, but prior ignorance is impossible

IP model:

$\mathcal{P}_{ignor}$: prior ignorance, but no consistency
(the probability interval is $[0, \frac{1}{2}]$ for all n)

$\mathcal{P}_{IDM(2)}$: consistency, but no prior ignorance
(the probability interval is $[0, \frac{1}{6}]$ when $n = 0$)

running example

Piatti, Zaffalon, and Trojani (2005) studied the behavior of the IDM model when the realization of each random variable $X_1, \ldots, X_n$ can be observed incorrectly with a known probability ε (the errors of observation are independent, conditional on the realizations of $X_1, \ldots, X_n$).

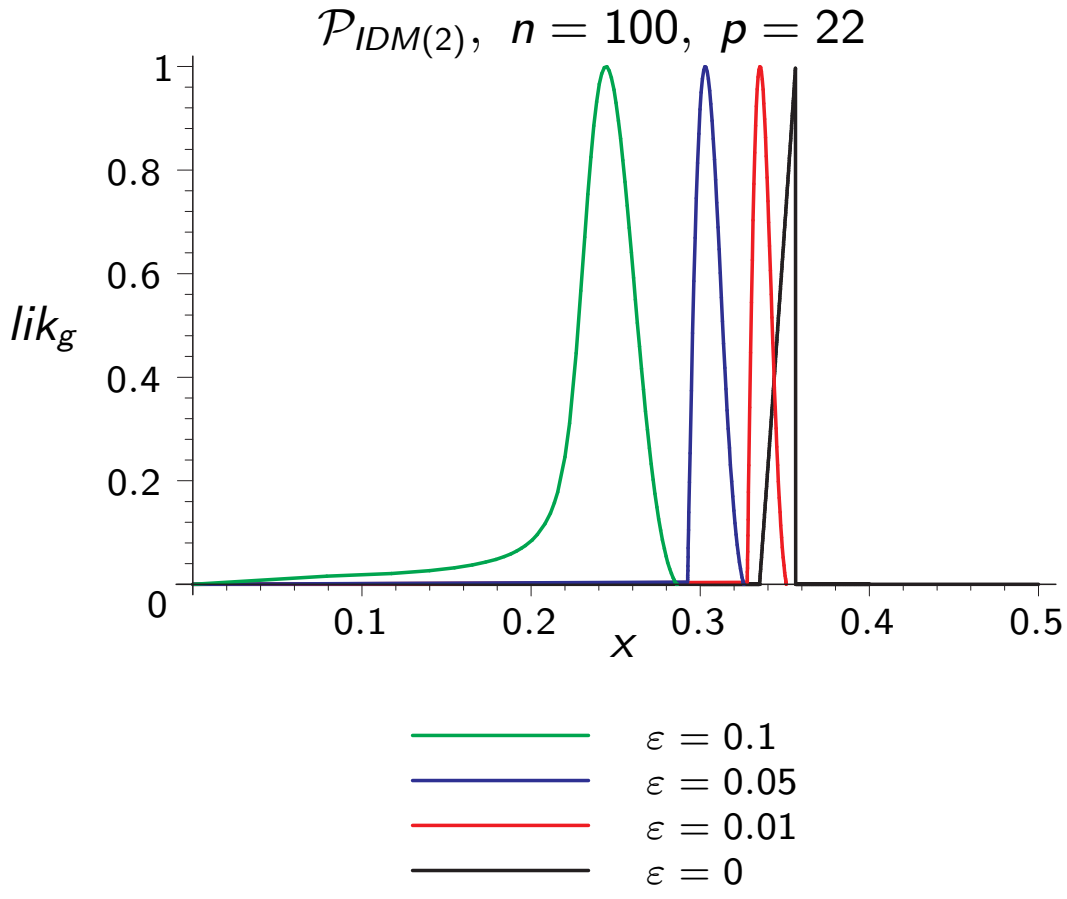
running example

Piatti, Zaffalon, and Trojani (2005) studied the behavior of the IDM model when the realization of each random variable $X_1, \ldots, X_n$ can be observed incorrectly with a known probability ε (the errors of observation are independent, conditional on the realizations of $X_1, \ldots, X_n$).

$g : P \mapsto P\{X_{n+1} + X_{n+2} = 1\}$

when $\varepsilon > 0$, the lower probability is 0 for all n

hence, for all $\varepsilon > 0$ the lower probability is **inconsistent**, and for all $n > 0$ it presents an important **discontinuity** when $\varepsilon \longrightarrow 0$



another simple example

The random objects $X \in \{a, b, c\}$ and $Y \in \{0, 1\}$ have the following joint probability distribution:

	$X = a$	$X = b$	$X = c$
$Y = 0$	0.01	0.01	0.70
$Y = 1$	0.04	0.04	0.20

$X \in \{b, c\}$ is observed: the conditional probability of $Y = 0$ is approximately 0.75

another simple example

The random objects $X \in \{a, b, c\}$ and $Y \in \{0, 1\}$ have the following joint probability distribution:

	$X = a$	$X = b$	$X = c$
$Y = 0$	0.01	0.01	0.70
$Y = 1$	0.04	0.04	0.20

$X \in \{b, c\}$ is observed: the conditional probability of $Y = 0$ is approximately 0.75

This probability value results from the assumption of "coarsening at random": more generally, De Cooman and Zaffalon (2004) assume that the observation O is a random subset of $\{a, b, c\}$, and consider the IP model described by the set $\mathcal{P}$ of all probabilistic models P such that the above holds, $P\{X = x,\ O = z\} = 0$ when $x \notin z$, and

$$P\{Y = 0 \,|\, X = x,\ O = z\} = P\{Y = 0 \,|\, X = x\} \quad \text{when } x \in z.$$

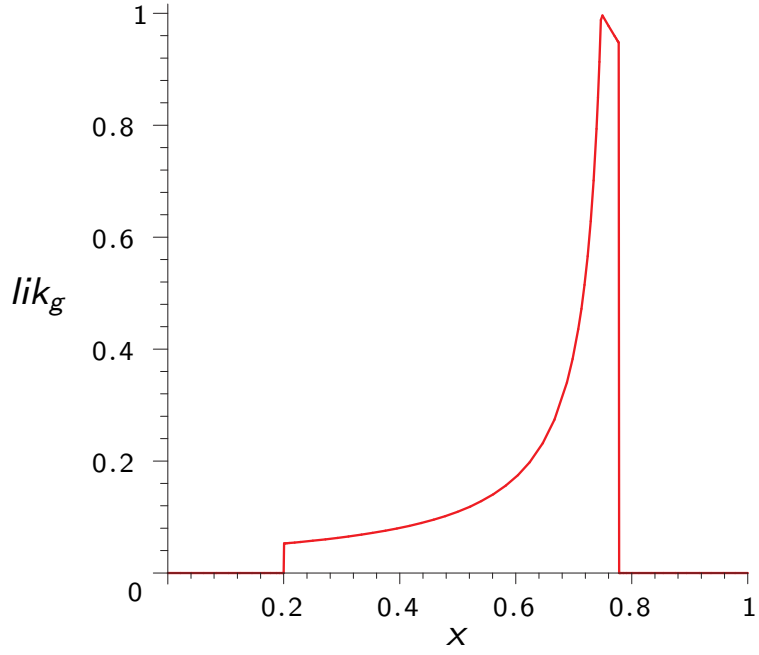
The probability interval of $Y = 0$ after having observed $O = \{b, c\}$ is approximately $[0.20, 0.78]$.

another simple example

$g : P \mapsto P\{Y = 0\}$

the maximum likelihood estimate corresponds to the probability value resulting from the assumption of "coarsening at random"

$supp(lik_g)$ corresponds to the probability interval resulting from the IP approach

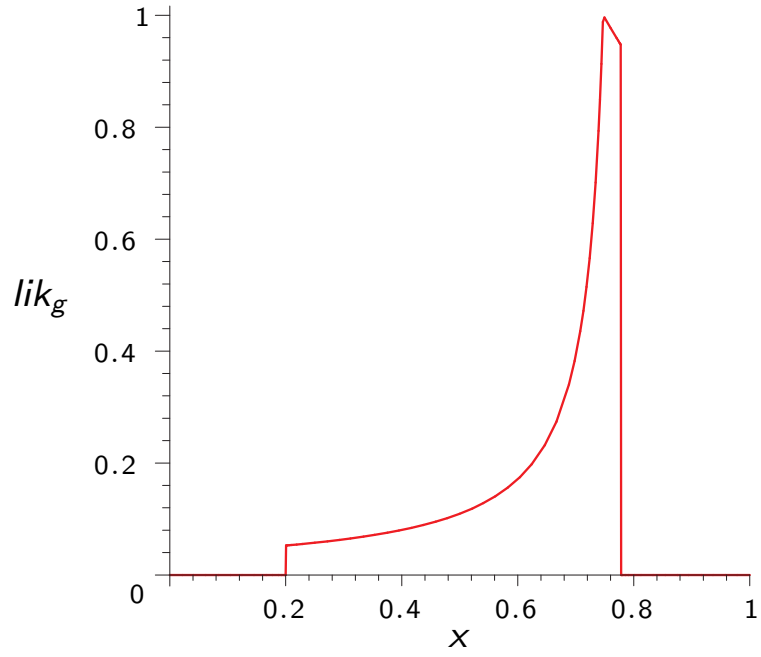


another simple example

$g : P \mapsto P\{Y = 0\}$

the maximum likelihood estimate corresponds to the probability value resulting from the assumption of "coarsening at random"

$supp(lik_g)$ corresponds to the probability interval resulting from the IP approach



This example can be made extreme with the following joint probability distribution, for a small $\varepsilon \in [0, 0.5]$:

	$X = a$	$X = b$	$X = c$
$Y = 0$	0	0	0.5
$Y = 1$	$0.5 - \varepsilon$	ε	0

The probability interval of $Y = 0$ after having observed $O = \{b, c\}$ is $[0, 1]$ for all $\varepsilon > 0$, but it collapses to the probability value 1 when $\varepsilon = 0$.

conclusion

A part of the information provided by the data is disregarded by the usual updating rule for IP models, and this leads to problems such as inconsistency or discontinuities.

conclusion

A part of the information provided by the data is disregarded by the usual updating rule for IP models, and this leads to problems such as inconsistency or discontinuities.

It does not seem possible to completely solve these problems in the framework of IP models; in particular, no updating rule whose results are always at least as precise as those of the usual updating rule can lead to statistical consistency (and sequential coherence would also be problematic).

conclusion

A part of the information provided by the data is disregarded by the usual updating rule for IP models, and this leads to problems such as inconsistency or discontinuities.

It does not seem possible to completely solve these problems in the framework of IP models; in particular, no updating rule whose results are always at least as precise as those of the usual updating rule can lead to statistical consistency (and sequential coherence would also be problematic).

To completely solve these problems, it seems necessary to store more information than it is possible in the framework of IP models: the hierarchical model provides a simple solution.

references

- Cattaneo, M. (2007). Statistical Decisions Based Directly on the Likelihood Function. PhD thesis, ETH Zurich. Available online at `e-collection.ethz.ch`.
- De Cooman, G., and Zaffalon, M. (2004). Updating beliefs with incomplete observations. Artif. Intell. 159.
- Hisdal, E. (1988). Are grades of membership probabilities? Fuzzy Sets Syst. 25.
- Pearson, K. (1920). The fundamental problem of practical statistics. Biometrika 13.
- Pelessoni, R., and Vicig, P. (2003). Convex imprecise previsions. Reliab. Comput. 9.
- Piatti, A., Zaffalon, M., and Trojani, F. (2005). Limits of learning from imperfect observations under prior ignorance: the case of the imprecise Dirichlet model. ISIPTA '05.