A hierarchical model based on the likelihood function

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

November 6, 2007

my research

- Master's thesis (ETH Zurich, March 2000):
  Inductive Inference and Nonmonotonic Reasoning

  $\longrightarrow$ ISIPTA '03:
  Combining belief functions issued from dependent sources

my research

- Master's thesis (ETH Zurich, March 2000):
  Inductive Inference and Nonmonotonic Reasoning

  $\rightarrow$ ISIPTA '03:
  Combining belief functions issued from dependent sources

- PhD thesis (ETH Zurich, March 2007):
  Statistical Decisions Based Directly on the Likelihood Function
  `http://e-collection.ethbib.ethz.ch`

  (ISIPTA '05: **Likelihood-based statistical decisions**)

my research

- Master's thesis (ETH Zurich, March 2000):
  Inductive Inference and Nonmonotonic Reasoning

  $\rightarrow$ ISIPTA '03:
  Combining belief functions issued from dependent sources

- PhD thesis (ETH Zurich, March 2007):
  Statistical Decisions Based Directly on the Likelihood Function
  `http://e-collection.ethbib.ethz.ch`

  (ISIPTA '05: **Likelihood-based statistical decisions**)

- Research fellowship from the "Stefano Franscini Fonds"
  (LMU Munich, October 2007 – September 2008):
  Decision making on the basis of a probabilistic-possibilistic hierarchical description of uncertain knowledge

the idea behind my thesis

- The likelihood function is central to statistics, and the most appreciated general methods of statistical inference are based directly on the likelihood function.

the idea behind my thesis

- The likelihood function is central to statistics, and the most appreciated general methods of statistical inference are based directly on the likelihood function.
- Wald unified statistics in his theory of decision functions, but the likelihood-based methods do not fit well into this perspective (they are in general suboptimal from the repeated sampling point of view).

the idea behind my thesis

- The likelihood function is central to statistics, and the most appreciated general methods of statistical inference are based directly on the likelihood function.
- Wald unified statistics in his theory of decision functions, but the likelihood-based methods do not fit well into this perspective (they are in general suboptimal from the repeated sampling point of view).
- In my thesis the decisions are based directly on the likelihood function, and the likelihood-based inference methods can be obtained as special cases.

the idea behind my thesis

- The likelihood function is central to statistics, and the most appreciated general methods of statistical inference are based directly on the likelihood function.
- Wald unified statistics in his theory of decision functions, but the likelihood-based methods do not fit well into this perspective (they are in general suboptimal from the repeated sampling point of view).
- In my thesis the decisions are based directly on the likelihood function, and the likelihood-based inference methods can be obtained as special cases.
- Through a new perspective on the relationships between likelihood-based methods, this approach suggests and justifies new methods based on the likelihood function.

the idea behind my thesis

- The likelihood function is central to statistics, and the most appreciated general methods of statistical inference are based directly on the likelihood function.
- Wald unified statistics in his theory of decision functions, but the likelihood-based methods do not fit well into this perspective (they are in general suboptimal from the repeated sampling point of view).
- In my thesis the decisions are based directly on the likelihood function, and the likelihood-based inference methods can be obtained as special cases.
- Through a new perspective on the relationships between likelihood-based methods, this approach suggests and justifies new methods based on the likelihood function.
- The resulting methods share the advantages of the likelihood-based inference methods: they are intuitive, generally applicable, conditional, dependent only on sufficient statistics, equivariant, parametrization invariant, asymptotically optimal (consistent) and efficient, and usually good from the repeated sampling point of view.

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 model of the reality under consideration; it assigns the probability $P(A)$ to the realization of the event $A \in \mathcal{A}$.

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 model of the reality under consideration; it assigns the probability $P(A)$ to the realization of the event $A \in \mathcal{A}$.

After having observed the event $A \in \mathcal{A}$, the **likelihood function** $lik : P \longmapsto P(A)$ on $\mathcal{P}$ describes the *relative* ability of the models to forecast the observed data.

The likelihood function can be interpreted as a measure of the *relative* plausibility of the models in the light of the observed data alone (proportional likelihood functions are equivalent).

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 model of the reality under consideration; it assigns the probability $P(A)$ to the realization of the event $A \in \mathcal{A}$.

After having observed the event $A \in \mathcal{A}$, the **likelihood function** $lik : P \longmapsto P(A)$ on $\mathcal{P}$ describes the *relative* ability of the models to forecast the observed data.

The likelihood function can be interpreted as a measure of the *relative* plausibility of the models in the light of the observed data alone (proportional likelihood functions are equivalent).

If we observe a second event $B \in \mathcal{A}$, then the likelihood of $P \in \mathcal{P}$ becomes $P(A \cap B) = P(A)\, P(B \mid A)$; that is, the new likelihood function is $lik\, lik'$, where $lik' : P \mapsto P(B \mid A)$.

the likelihood ratio

The **likelihood ratio** test discards the hypothesis $\mathcal{H} \subseteq \mathcal{P}$ when

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

is sufficiently small (where $\frac{1}{c} = \sup_{P \in \mathcal{P}} lik(P)$).

In regular problems, under the hypothesis $\mathcal{H}$, the statistic $-2 \log LR(\mathcal{H})$ is asymptotically χ^2 distributed (the number of degrees of freedom is the difference in dimensionality between $\mathcal{P}$ and $\mathcal{H}$).

the likelihood ratio

The **likelihood ratio** test discards the hypothesis $\mathcal{H} \subseteq \mathcal{P}$ when

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

is sufficiently small (where $\frac{1}{c} = \sup_{P \in \mathcal{P}} lik(P)$).

In regular problems, under the hypothesis $\mathcal{H}$, the statistic $-2 \log LR(\mathcal{H})$ is asymptotically χ^2 distributed (the number of degrees of freedom is the difference in dimensionality between $\mathcal{P}$ and $\mathcal{H}$).

The (nonadditive) measure $LR : 2^{\mathcal{P}} \rightarrow [0,1]$ is normalized and completely maxitive, that is:

$$LR(\mathcal{P}) = 1 \qquad \text{and} \qquad LR\left(\bigcup_{\mathcal{H} \in \mathcal{S}} \mathcal{H}\right) = \sup_{\mathcal{H} \in \mathcal{S}} LR(\mathcal{H}) \quad \text{for all } \mathcal{S} \subseteq 2^{\mathcal{P}}.$$

A completely maxitive measure LR on a set $\mathcal{P}$ is determined by its density function $LR^{\downarrow} : P \mapsto LR\{P\}$ on $\mathcal{P}$, since $LR(\mathcal{H}) = \sup_{P \in \mathcal{H}} LR^{\downarrow}(P)$.

the hierarchical model

By interpreting the likelihood function as a measure of the relative plausibility of the models in $\mathcal{P}$, we obtain a probabilistic-possibilistic hierarchical description of uncertain knowledge:

$$\overset{LR}{\mathcal{P}}.$$

the hierarchical model

By interpreting the likelihood function as a measure of the relative plausibility of the models in $\mathcal{P}$, we obtain a probabilistic-possibilistic hierarchical description of uncertain knowledge:

$$\overset{LR}{\mathcal{P}}.$$

The uncertain knowledge about the value of a function $g : \mathcal{P} \mapsto \mathcal{G}$ is described by the (relative) possibility measure $LR \circ g^{-1}$ induced on $\mathcal{G}$.

the hierarchical model

By interpreting the likelihood function as a measure of the relative plausibility of the models in $\mathcal{P}$, we obtain a probabilistic-possibilistic hierarchical description of uncertain knowledge:

$$\overset{LR}{\mathcal{P}}.$$

The uncertain knowledge about the value of a function $g : \mathcal{P} \mapsto \mathcal{G}$ is described by the (relative) possibility measure $LR \circ g^{-1}$ induced on $\mathcal{G}$.

When $C \in \mathcal{A}$ is observed, the hierarchical model $\overset{LR}{\mathcal{P}}$ is updated to $\overset{LR'}{\mathcal{P}'}$, where $\mathcal{P}' = \{P(\cdot \mid C) : P \in \mathcal{P},\ P(C) > 0\}$ and

$$LR'\{P'\} \propto \sup\left\{LR\{P\}\, P(C) : P \in \mathcal{P},\ P(\cdot \mid C) = P'\right\} \quad \text{for all } P' \in \mathcal{P}'.$$

the description of ignorance

A constant likelihood function describes the complete absence of information for discrimination between the models in $\mathcal{P}$: in this case LR is denoted by $\varnothing$ (that is, $\varnothing^{\downarrow} = 1$).

the description of ignorance

A constant likelihood function describes the complete absence of information for discrimination between the models in $\mathcal{P}$: in this case LR is denoted by $\varnothing$ (that is, $\varnothing^{\downarrow} = 1$).

When we use no prior information about the relative plausibility of the elements of $\mathcal{P}$, we start with the hierarchical model $\overset{\varnothing}{\mathcal{P}}$; but we can also start with the hierarchical model $\overset{LR}{\mathcal{P}}$ for some prior non-constant likelihood function $LR^{\downarrow}$ on $\mathcal{P}$, interpreted as the likelihood function induced by the prior information.

the description of ignorance

A constant likelihood function describes the complete absence of information for discrimination between the models in $\mathcal{P}$: in this case LR is denoted by $\varnothing$ (that is, $\varnothing^{\downarrow} = 1$).

When we use no prior information about the relative plausibility of the elements of $\mathcal{P}$, we start with the hierarchical model $\overset{\varnothing}{\mathcal{P}}$; but we can also start with the hierarchical model $\overset{LR}{\mathcal{P}}$ for some prior non-constant likelihood function $LR^{\downarrow}$ on $\mathcal{P}$, interpreted as the likelihood function induced by the prior information.

The fundamental qualitative difference between a probability measure π on $\mathcal{P}$ and a possibility measure LR on $\mathcal{P}$ is that when $\mathcal{H}$ and $\mathcal{H}'$ are two (measurable) disjoint subsets of $\mathcal{P}$,

$$\pi(\mathcal{H}') > 0 \quad \Rightarrow \quad \pi(\mathcal{H} \cup \mathcal{H}') > \pi(\mathcal{H}),$$

while $\quad LR(\mathcal{H}') > 0 \quad$ and $\quad LR(\mathcal{H} \cup \mathcal{H}') = LR(\mathcal{H}) \quad$ are compatible.

the usual approaches to statistics

- In the classical approach we use a model of the form $\overset{\varnothing}{\mathcal{P}}$, and we never update it (we base the conclusions on expected values).

the usual approaches to statistics

- In the classical approach we use a model of the form $\overset{\varnothing}{\mathcal{P}}$, and we never update it (we base the conclusions on expected values).

- In the Bayesian approach we use a model of the form $\{\overset{\varnothing}{P}\}$, and we update it by conditioning P.

the usual approaches to statistics

- In the classical approach we use a model of the form $\overset{\varnothing}{\mathcal{P}}$, and we never update it (we base the conclusions on expected values).

- In the Bayesian approach we use a model of the form $\{\overset{\varnothing}{P}\}$, and we update it by conditioning P.

- In the IP approach we use a model of the form $\overset{\varnothing}{\mathcal{P}}$, but when we update it (by means of "regular extension") we throw away the information contained in the likelihood function.

the usual approaches to statistics

- In the classical approach we use a model of the form $\overset{\varnothing}{\mathcal{P}}$, and we never update it (we base the conclusions on expected values).

- In the Bayesian approach we use a model of the form $\{\overset{\varnothing}{P}\}$, and we update it by conditioning P.

- In the IP approach we use a model of the form $\overset{\varnothing}{\mathcal{P}}$, but when we update it (by means of "regular extension") we throw away the information contained in the likelihood function.

  If the elements of $\mathcal{P}$ represent the opinions of a group of Bayesian experts, then the updating by means of "regular extension" 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.

"regular extension" leads to inconsistency

An example by Wilson (ISIPTA '01):

Let $P(Y=0) = P(Y=1) = 0.5$,
and let $X_1, X_2, \ldots, X_{100} \in \{0,1\}$ be i.i.d. conditional on Y
with $\quad P(X_i = 1 \mid Y = 0) = 0.5 \quad$ and $\quad 0.1 \leq P(X_i = 1 \mid Y = 1) \leq 0.6.$

“regular extension” leads to inconsistency

An example by Wilson (ISIPTA '01):

Let $P(Y = 0) = P(Y = 1) = 0.5$,
and let $X_1, X_2, \ldots, X_{100} \in \{0, 1\}$ be i.i.d. conditional on Y
with $\quad P(X_i = 1 \mid Y = 0) = 0.5 \quad$ and $\quad 0.1 \leq P(X_i = 1 \mid Y = 1) \leq 0.6$.

After having observed the realizations of $X_1, X_2, \ldots, X_{100}$ with mean 0.2, we would expect the conditional distribution of Y to be concentrated on 1 (since $Y = 1$ is compatible with the observations, while $Y = 0$ is not), but when we update the model by means of “regular extension”, we almost obtain complete ignorance about the value of Y (the posterior interval probability of $Y = 0$ is approximately [0.000000004, 0.999999]).

“regular extension” leads to inconsistency

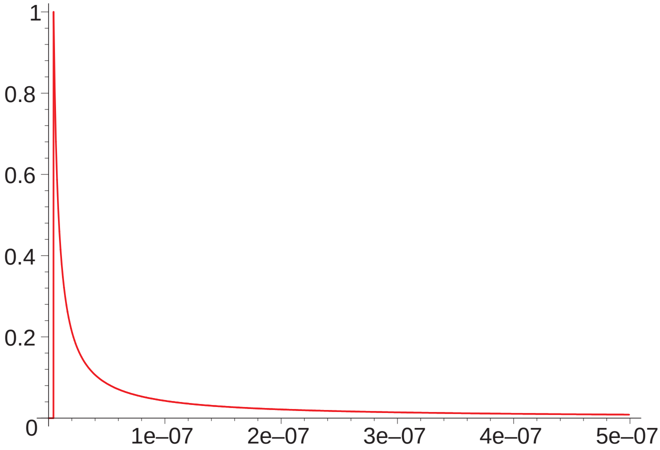
An example by Wilson (ISIPTA '01):

Let $P(Y = 0) = P(Y = 1) = 0.5$,
and let $X_1, X_2, \ldots, X_{100} \in \{0, 1\}$ be i.i.d. conditional on Y
with $P(X_i = 1 \mid Y = 0) = 0.5$ and $0.1 \leq P(X_i = 1 \mid Y = 1) \leq 0.6$.

After having observed the realizations of $X_1, X_2, \ldots, X_{100}$ with mean 0.2, we would expect the conditional distribution of Y to be concentrated on 1 (since $Y = 1$ is compatible with the observations, while $Y = 0$ is not), but when we update the model by means of “regular extension”, we almost obtain complete ignorance about the value of Y (the posterior interval probability of $Y = 0$ is approximately [0.000000004, 0.999999]).

This interval probability is the support of the density function $(LR \circ g^{-1})^{\downarrow}$ of the conditional “fuzzy probability” that we obtain when we consider the likelihood function on the set $\mathcal{P}$ of the models P and we define $g(P)$ as the conditional probability of $Y = 0$ under the model P; but the conditional “fuzzy probability” of $Y = 0$ is concentrated toward 0, in agreement with the intuition that the conditional distribution of Y should be concentrated on 1.

the conditional "fuzzy probability" of $Y = 0$



another simple example

Assume that we have the following joint probability distribution P for the random variables X and Y:

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

another simple example

Assume that we have the following joint probability distribution P for the random variables X and Y:

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

We observe $X \in \{b, c\}$, and $\quad P(Y = 0 \mid X \in \{b, c\}) \approx 0.75$.

another simple example

Assume that we have the following joint probability distribution P for the random variables X and Y:

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

We observe $X \in \{b, c\}$, and $\quad P(Y = 0 \,|\, X \in \{b, c\}) \approx 0.75$.

This conclusion corresponds to 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 the probability measure P satisfies $\quad P(X = x,\ O = z) = 0 \quad$ when $x \notin z$, and

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

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

the conditional probability of $Y = 0$