Fuzzy Probabilities Based on the Likelihood Function

Marco E.G.V. Cattaneo

Institut für Statistik, Ludwig-Maximilians-Universität München, München, Germany

Abstract. If we interpret the statistical likelihood function as a measure of the relative plausibility of the probabilistic models considered, then we obtain a hierarchical description of uncertain knowledge, offering a unified approach to the combination of probabilistic and possibilistic uncertainty. The fundamental advantage of the resulting fuzzy probabilities with respect to imprecise probabilities is the ability of using all the information provided by the data.

Keywords: Likelihood function, Hierarchical model, Fuzzy probabilities, Imprecise probabilities, Statistical inconsistency.

1 Introduction

This paper presents a probabilistic-possibilistic hierarchical model based on the likelihood function. Thanks to the intuitivity and asymptotic properties of the likelihood function, the hierarchical model is an ideal basis for inference and decision making: this aspect is analyzed in [2]. The hierarchical model can be interpreted as a fuzzy probability measure, and offers a unified approach to the combination of probabilistic and possibilistic uncertainty.

Fuzzy probabilities generalize imprecise probabilities by additionally considering the relative plausibility of different values in the probability intervals (imprecise probabilities correspond to the crisp case of fuzzy probabilities). By abandoning the crispness of imprecise probabilities, the hierarchical model solves a fundamental problem of the imprecise probability approach: its statistical inconsistency.

2 Hierarchical Model

Let $\mathscr{P}$ be a set of probability measures on a measurable space $(\Omega, \mathscr{A})$ such that $\mathscr{A}$ contains all singletons of Ω. Each $P \in \mathscr{P}$ is interpreted as a probabilistic model of the reality under consideration. The interpretation of probability is not important: for instance the elements of $\mathscr{P}$ can be statistical models, or describe the forecasts of a group of experts.

When an event $A \in \mathscr{A}$ is observed, the *likelihood function*

$$lik : P \longmapsto P(A)$$

on $\mathscr{P}$ describes the relative ability of the probabilistic models in $\mathscr{P}$ to forecast the observed data. Spurious modifications of the situation considered or of its mathematical

D. Dubois et al. (Eds.): Soft Methods for Hand. Var. and Imprecision, ASC 48, pp. 43–50, 2008.
springerlink.com

representation can lead to likelihood functions proportional to lik. Therefore, proportional likelihood functions are considered equivalent; in fact, Fisher [8] defined the likelihood of a statistical model as a quantity *proportional* to the probability of the observed data. Hence, only ratios $lik(P)/lik(P')$ of the values of lik for different $P, P' \in \mathscr{P}$ have meaning: Kullback and Leibler [11] interpreted $\log[lik(P)/lik(P')]$ as the information in A for discrimination in favor of P against P'. When the realization of a continuous random object is observed, the usual definition of likelihood function in terms of density can be seen as an approximation of lik (see [2, Section 1.2]).

The likelihood function can thus be interpreted as a measure of the relative plausibility of the probabilistic models in the light of the observed data alone. Under each probabilistic model $P \in \mathscr{P}$, the likelihood ratio $lik(P)/lik(P')$ of P against a different probabilistic model $P' \in \mathscr{P}$ almost surely increases without bound when more and more data are observed, and consequently lik tends to concentrate around P, if some regularity conditions are satisfied. Thanks to this asymptotic property and to its intuitivity, the likelihood function is an ideal basis for statistical inference and decision making (see [13] for an introduction to the likelihood approach to statistics).

Example 1. Let $\mathscr{P} = \{P_p : p \in [0.1, 0.6]\}$ be a set of probability measures on a measurable space $(\Omega, \mathscr{A})$, such that for each $P_p \in \mathscr{P}$ the random variables $X_0, \ldots, X_{100} : \Omega \to \{0,1\}$ satisfy the following conditions: $P_p\{X_0 = 0\} = \frac{1}{2}$, and conditional on the realization of X_0 the random variables $X_1, \ldots, X_{100}$ are independent with $P_p\{X_i = 1 \,|\, X_0 = 0\} = \frac{1}{2}$ and $P_p\{X_i = 1 \,|\, X_0 = 1\} = p$ for all $i \in \{1, \ldots, 100\}$.

The realizations of $X_1, \ldots, X_{100}$ are observed: 20 of them take the value 1. The resulting likelihood function

$$lik : P_p \longmapsto \tfrac{1}{2}\left(\tfrac{1}{2}\right)^{100} + \tfrac{1}{2}\, p^{20}\,(1-p)^{80}$$

on $\mathscr{P}$ is concentrated around $P_{0.2}$, which is the most plausible element of $\mathscr{P}$ in the light of the observed data alone. The case with $X_0 = 0$ has almost no influence on the form of the likelihood function, and in fact this case is extremely implausible in the light of the observed data and of the probabilistic models considered.

The likelihood function lik measures the relative plausibility of the elements of $\mathscr{P}$, but a measure of the relative plausibility of the subsets of $\mathscr{P}$ is often needed. A simple and effective way to obtain it consists in defining the plausibility of a set of probabilistic models as the plausibility of its best element: the result is the set function

$$\mathscr{H} \longmapsto \sup_{P \in \mathscr{H}} lik(P)$$

on the power set $2^{\mathscr{P}}$ of $\mathscr{P}$ (in this paper, $\sup \varnothing = 0$). Proportional set functions of this form are equivalent, since they correspond to equivalent likelihood functions: to underline this relative meaning, the expression “relative plausibility measure” is used in [2] to denote an equivalence class of proportional set functions of this form. Their normalized version LR associates to each $\mathscr{H} \subseteq \mathscr{P}$ the corresponding likelihood ratio statistic

$$LR(\mathscr{H}) = \frac{\sup_{P \in \mathscr{H}} lik(P)}{\sup_{P \in \mathscr{P}} lik(P)}.$$

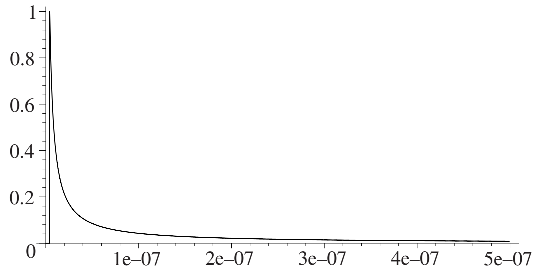

Fig. 1. Profile likelihood function from Example 2 and membership function of fuzzy probability from Example 3

The *likelihood ratio test* discards the hypothesis that the data were generated by some $P \in \mathscr{H}$ if $LR(\mathscr{H})$ is sufficiently small.

Let $g : \mathscr{P} \to \mathscr{G}$ be a function. The likelihood function lik on $\mathscr{P}$ induces the (normalized) *profile likelihood function*

$$lik_g : \gamma \longmapsto LR(g^{-1}\{\gamma\}) \propto \sup_{P \in \mathscr{P} : g(P) = \gamma} lik(P)$$

on $\mathscr{G}$ (in this paper, g^{-1} denotes the set function associating to each subset of $\mathscr{G}$ its inverse image under g). The profile likelihood function lik_g measures the relative plausibility of the values of g, on the basis of the above definition of plausibility for a set of probabilistic models. The *maximum likelihood estimate* $\hat{\gamma}_{ML}$ of $g(P)$ is the $\gamma \in \mathscr{G}$ maximizing $lik_g(\gamma)$ (that is, $lik_g(\hat{\gamma}_{ML}) = 1$), when such a γ exists and is unique. The *likelihood-based confidence region* for $g(P)$ with cutoff point $\alpha \in (0,1)$ is the set $\{\gamma \in \mathscr{G} : lik_g(\gamma) > \alpha\}$: it is the smallest $G \subseteq \mathscr{G}$ such that $LR\{P \in \mathscr{P} : g(P) \notin G\} \leq \alpha$.

Example 2. Consider the situation of Example 1, and let $g : \mathscr{P} \to [0,1]$ associate to each probabilistic model in $\mathscr{P}$ the probability of $X_0 = 0$ conditional on the observed realizations of $X_1, \ldots, X_{100}$:

$$g : P_p \longmapsto \frac{\left(\frac{1}{2}\right)^{100}}{\left(\frac{1}{2}\right)^{100} + p^{20}\,(1-p)^{80}}.$$

Figure 1 shows the graph of the profile likelihood function lik_g on $[0, 5 \cdot 10^{-7}]$: as expected, lik_g is extremely concentrated near 0, because $X_0 = 1$ is compatible with the observed data, while $X_0 = 0$ is not. In fact, the maximum likelihood estimate of $g(P_p)$ is $\hat{\gamma}_{ML} \approx 0.04 \cdot 10^{-7}$, and the likelihood-based confidence region for $g(P_p)$ with cutoff point $\alpha = 0.01$ corresponds approximately to the interval $(0.04 \cdot 10^{-7}, 4.26 \cdot 10^{-7})$.

The probabilistic models in $\mathscr{P}$ and the likelihood function lik on $\mathscr{P}$ can be interpreted as the two levels of a *hierarchical model* of the reality under consideration. The two levels describe different kinds of uncertain knowledge: in the first level the uncertainty is stochastic, while in the second one it is about which of the probabilistic models in $\mathscr{P}$ is the best representation of the reality. It is important to underline that no probabilistic model in $\mathscr{P}$ is assumed to be in some sense "true": the elements of $\mathscr{P}$ are simply interpreted as more or less plausible representations of the reality (this interpretation of the hierarchical model is shared by Edwards [7]). By contrast, the use of a probability measure on $\mathscr{P}$, suggested by the Bayesian approach, carries the implicit assumption that exactly one of the probabilistic models in $\mathscr{P}$ is "true" (see [2, Section 3.1]).

The definition of likelihood function implies that when an event $A \in \mathscr{A}$ is observed, the two levels $\mathscr{P}$ and lik of the hierarchical model are updated to

$$\mathscr{P}' = \{P(\cdot\,|A) : P \in \mathscr{P}, P(A) > 0\} \tag{1}$$

$$\text{and to} \quad lik' : P' \longmapsto \sup_{P \in \mathscr{P} : P(\cdot\,|A) = P'} lik(P)\,P(A),$$

respectively. When A is the first observed event, the *prior* likelihood function lik can be interpreted as a (subjective) measure of the relative plausibility of the probabilistic models in $\mathscr{P}$ according to the prior information. The choice of a prior likelihood function on $\mathscr{P}$ seems to be better supported by intuition than the choice of a prior probability measure on $\mathscr{P}$: in particular, a constant likelihood function describes *complete ignorance* (in the sense of absence of information for discrimination between the probabilistic models). In fact, if lik is constant, then lik' is proportional to the profile likelihood function on $\mathscr{P}'$ induced by the observation A and the conditioning $P \mapsto P(\cdot\,|A)$. Moreover, the choice of a prior likelihood function can be based on analogies with the likelihood functions induced by hypothetical data (see also [3]).

3 Fuzzy Probabilities

A *possibility distribution* on a set $\mathscr{G}$ is a function $\pi : \mathscr{G} \to [0,1]$. The possibility measure on $\mathscr{G}$ with possibility distribution π is the set function

$$G \longmapsto \sup_{\gamma \in G} \pi(\gamma)$$

on $2^{\mathscr{G}}$. A possibility distribution π on $\mathscr{G}$ can also be considered as the *membership function* of a fuzzy subset of $\mathscr{G}$ (see [17]); when π is *crisp* (that is, π can take only the values 0 and 1), the subset is not fuzzy and π is its indicator function on $\mathscr{G}$. The likelihood ratio statistic LR is a possibility measure on $\mathscr{P}$ with possibility distribution proportional to the likelihood function lik on $\mathscr{P}$. In fact, the membership function of a fuzzy set has often been interpreted as a likelihood function (see for example [10, 5]), even though proportional membership functions were not always considered equivalent (see for instance [6]). In the present paper, membership functions and possibility distributions are interpreted as *proportional* to likelihood functions. Hence, it suffices to consider normalized fuzzy sets and normalized possibility measures (that is, $\sup_{\gamma \in \mathscr{G}} \pi(\gamma) = 1$ is assumed), but grades of membership and degrees of possibility have only a relative meaning.

The hierarchical model considered in the previous section can thus be interpreted as consisting of a probabilistic level (described by $\mathscr{P}$) and a possibilistic level (described by LR). That is, it can be interpreted as a probabilistic-possibilistic hierarchical description of uncertain knowledge about $\omega \in \Omega$. Both the purely probabilistic and the purely possibilistic descriptions of uncertain knowledge about $\omega \in \Omega$ appear as special cases. In fact, when $\mathscr{P}$ is a singleton, the uncertainty about $\omega \in \Omega$ is purely probabilistic (*LR* on $\mathscr{P} = \{P\}$ contains no information, since its meaning is only relative). By contrast, when $\mathscr{P}$ consists of all the Dirac measures (that is, $\mathscr{P} = \{\delta_\omega : \omega \in \Omega\}$ with $\delta_\omega\{\omega\} = 1$), the uncertainty about $\omega \in \Omega$ is purely possibilistic (*LR* can be considered as a possibility measure on Ω, since each $\delta_\omega \in \mathscr{P}$ can be identified with the corresponding $\omega \in \Omega$).

The hierarchical model can also be interpreted as a fuzzy probability measure on $(\Omega, \mathscr{A})$, in the sense that it is a fuzzy subset of the set of all probability measures on $(\Omega, \mathscr{A})$, with membership function proportional to *lik* on $\mathscr{P}$ and constant equal to 0 outside $\mathscr{P}$. More generally, the uncertain knowledge about the value $g(P)$ of a function $g : \mathscr{P} \to \mathscr{G}$ is described by the induced possibility measure $LR \circ g^{-1}$ on $\mathscr{G}$; that is, by the fuzzy subset of $\mathscr{G}$ with membership function lik_g. In particular, when $g : \mathscr{P} \to \mathbb{R}$, the uncertain knowledge about $g(P)$ is described by a fuzzy number (that is, a fuzzy subset of $\mathbb{R}$). For example, g can associate to each probabilistic model P the expectation $g(P) = E_P(X)$ of a random variable X, or the probability $g(P) = P(A)$ of an event $A \in \mathscr{A}$: the membership function lik_g describes then the *fuzzy expectation* of X, or the *fuzzy probability* of A, respectively. Sometimes a fuzzy number can be a satisfactory conclusion about the value of $g(P)$, but it is often necessary to evaluate the fuzzy number by a single real number (such as the maximum likelihood estimate $\hat{\gamma}_{ML}$) or by a couple of real numbers (such as the infimum and the supremum of a likelihood-based confidence region $\{\gamma \in \mathbb{R} : lik_g(\gamma) > \alpha\}$). The discussion on how to evaluate a fuzzy number by one or more real numbers goes beyond the scope of the present paper, but see [2, Section 4.1] for some interesting results (to each evaluation method corresponds a likelihood-based decision criterion).

Example 3. The prior fuzzy probability measure on $(\Omega, \mathscr{A})$ considered in Examples 1 and 2 is crisp, in the sense that its membership function on the set of all probability measures on $(\Omega, \mathscr{A})$ is crisp. In fact, the only prior (non-stochastic) uncertainty is about the value of the probability of $X_i = 1$ conditional on $X_0 = 1$ (with $i \in \{1, \ldots, 100\}$), and the only prior information about this value is that it lies in the interval $[0.1, 0.6]$. But the updated fuzzy probability measure on $(\Omega, \mathscr{A})$ obtained after having observed the realizations of $X_1, \ldots, X_{100}$ is not crisp anymore: the fuzzy (conditional) probability of $X_0 = 0$ has membership function lik_g (plotted in Figure 1). Hence, any reasonable evaluation of the fuzzy (conditional) probability of $X_0 = 0$ by a real number (such as the maximum likelihood estimate $\hat{\gamma}_{ML} \approx 0.04 \cdot 10^{-7}$, or the lower and upper evaluations $0.04 \cdot 10^{-7}$ and $4.26 \cdot 10^{-7}$ considered at the end of Example 2) would be approximately 0.

The hierarchical model offers a unified approach to the combination of probabilistic and possibilistic uncertainty (in particular, fuzzy data would pose no problem). Since membership functions and possibility distributions are interpreted as proportional to likelihood functions, the rules for manipulating fuzzy probabilities are implied by the well-established theories of probability and likelihood (the same holds for the approach of De Cooman [4], which uses a different interpretation of possibility measures). By contrast, approaches based on the arithmetic of fuzzy numbers (see for example [14, 1]) face the problem of choosing and justifying such rules: the choice of a consistent way of updating the fuzzy probability models in the light of data seems to be particularly difficult.

4 Imprecise Probabilities

The mathematical representations of reality used in the classical and Bayesian approaches to statistics can be considered as special cases of the hierarchical model (see [2, Section 3.2]). By contrast, the imprecise probability model cannot be considered as a special case of the hierarchical model, because the updating rules are different. The mathematical representation of reality used in the imprecise probability approach to statistics can be described as a (convex) set $\mathscr{P}$ of probabilistic models, without information for discrimination between them. This corresponds to a hierarchical model with constant likelihood function on $\mathscr{P}$, but the imprecise probability model is usually updated by *regular extension* (see [15, Appendix J]): that is, by conditioning each $P \in \mathscr{P}$ on the observed data, without considering the information provided by the likelihood function on $\mathscr{P}$. More precisely, when an event $A \in \mathscr{A}$ is observed, the set $\mathscr{P}$ is updated to the set $\mathscr{P}'$ as in (1), but the constant likelihood function on $\mathscr{P}$ is not updated: the likelihood function on $\mathscr{P}'$ is still constant; that is, the information in A for discrimination between the elements of $\mathscr{P}$ is disregarded.

For instance, if the probabilistic models in $\mathscr{P}$ describe the opinions of a group of Bayesian experts, then the updating by 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. This is not very reasonable, and in fact the updating by regular extension can lead to *inconsistency*, in the statistical sense of not tending to the correct conclusion, even when the amount of information provided by the data tends to infinity.

Example 4. The set $\mathscr{P}$ of probabilistic models considered in Examples 1, 2, and 3 can be interpreted as an imprecise probability measure on $(\Omega, \mathscr{A})$. If it is updated by regular extension, when the realizations of $X_1, \ldots, X_{100}$ are observed, then the resulting imprecise probability measure is described by the set $\mathscr{P}'$. In particular, the resulting uncertain knowledge about the (conditional) probability of $X_0 = 0$ is described by the lower and upper probabilities

$$\inf_{P' \in \mathscr{P}'} P'\{X_0 = 0\} \approx 4.26 \cdot 10^{-9} \quad \text{and} \quad \sup_{P' \in \mathscr{P}'} P'\{X_0 = 0\} \approx 1 - 6.77 \cdot 10^{-7}.$$

That is, despite the overwhelming information in favor of $X_0 = 1$ against $X_0 = 0$, almost complete ignorance about the (conditional) probabilities of $X_0 = 0$ and $X_0 = 1$ is obtained when the imprecise probability model is updated by regular extension (it is important to note that these results do not change when $\mathscr{P}$ is replaced by its convex hull). In fact, the resulting interval probability of $X_0 = 0$ is the support $\{\gamma \in [0,1] : lik_g(\gamma) > 0\}$ of the membership function lik_g of the fuzzy (conditional) probability of $X_0 = 0$ (plotted in Figure 1): lik_g is extremely concentrated near 0, but this information is disregarded when updating the imprecise probability model by regular extension (the present example was proposed by Wilson [16]).

The imprecise probability model can be seen as the crisp (and convex) case of the fuzzy probability model, but in general the crispness of the fuzzy probability model is lost when it is updated. Hence, from the point of view of the hierarchical model, the regular extension forces the crispness of the updated model by disregarding a part of the information provided by the data, and this can lead to statistical inconsistency. Many authors (see for example [16, 12]) have replaced, in particular problems, the regular extension with alternative updating rules making use of some information contained in the likelihood function on $\mathscr{P}$. But no alternative rule updating $\mathscr{P}$ to a subset of $\mathscr{P}'$ can assure the statistical consistency, because any discarded probabilistic model can become the most plausible one in the light of new data.

5 Conclusion

Statistical inconsistency is a fundamental problem of the theory of imprecise probabilities: a simple solution is to generalize imprecise probabilities to fuzzy probabilities, and use the probabilistic-possibilistic hierarchical model presented in this paper. In fact, fuzzy probabilities seem to be very intuitive: many authors (see for example [9, 4]) have studied models similar to the hierarchical one to accommodate the fact that usually not all the values in probability intervals are considered equally plausible.

References

1. Buckley, J.J.: Fuzzy Probability and Statistics. Studies in Fuzziness and Soft Computing, vol. 196. Springer, New York (2006)
2. Cattaneo, M.: Statistical Decisions Based Directly on the Likelihood Function. PhD Thesis, ETH Zurich (2007), `http://e-collection.ethz.ch`
3. Dahl, F.A.: Representing human uncertainty by subjective likelihood estimates. Internat. J. Approx. Reason 39(1), 85–95 (2005)
4. De Cooman, G.: A behavioural model for vague probability assessments. Fuzzy Sets Syst. 154, 305–358 (2005)
5. Dubois, D.: Possibility theory and statistical reasoning. Comput. Stat. Data Anal. 51, 47–69 (2006)
6. Dubois, D., Moral, S., Prade, H.: A semantics for possibility theory based on likelihoods. J. Math. Anal. Appl. 205(2), 359–380 (1997)
7. Edwards, A.W.F.: Likelihood, 2nd edn. Johns Hopkins University Press, Baltimore (1992)
8. Fisher, R.A.: On the mathematical foundations of theoretical statistics. Philos. Trans. R. Soc. Lond., Ser. A 222, 309–368 (1922)
9. Gärdenfors, P., Sahlin, N.E.: Unreliable probabilities, risk taking, and decision making. Synthese 53, 361–386 (1982)
10. Hisdal, E.: Are grades of membership probabilities? Fuzzy Sets Syst. 25, 325–348 (1988)
11. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. 22, 79–86 (1951)
12. Moral, S.: Calculating uncertainty intervals from conditional convex sets of probabilities. In: Dubois, D., Wellman, M.P. (eds.) Proceedings of Uncertainty in Artificial Intelligence (UAI 1992, Stanford, CA, USA), San Mateo, CA, USA, pp. 199–206. Morgan Kaufmann, San Francisco (1992)

13. Pawitan, Y.: In All Likelihood: Statistical Modelling and Inference Using Likelihood. Oxford University Press, New York (2001)
14. Viertl, R., Hareter, D.: Beschreibung und Analyse unscharfer Information. Springer, Wien (2006)
15. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Monographs on Statistics and Applied Probability, vol. 42. Chapman and Hall, Ltd., London (1991)
16. Wilson, N.: Modified upper and lower probabilities based on imprecise likelihoods. In: De Cooman, G., Fine, T.L., Seidenfeld, T. (eds.) Proceedings of the 2nd International Symposium on Imprecise Probabilities and Their Applications (ISIPTA 2001, Ithaca, USA), pp. 370–378. Shaker Publishing, Maastricht (2001)
17. Zadeh, L.A.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst. 1(1), 3–28 (1978)