Statistical Science
2021, Vol. 36, No. 2, 205–209
https://doi.org/10.1214/21-STS765D
Main article: https://doi.org/10.1214/19-STS765

Comment: On Focusing, Soft and Strong Revision of Choquet Capacities and Their Role in Statistics

Thomas Augustin and Georg Schollmeyer

Abstract. We congratulate Ruobin Gong and Xiao-Li Meng on their thought-provoking paper demonstrating the power of imprecise probabilities in statistics. In particular, Gong and Meng clarify important statistical paradoxes by discussing them in the framework of generalized uncertainty quantification and different conditioning rules used for updating. In this note, we characterize all three conditioning rules as envelopes of certain sets of conditional probabilities. This view also suggests some generalizations that can be seen as compromise rules. Similar to Gong and Meng, our derivations mainly focus on Choquet capacities of order 2, and so we also briefly discuss in general their role as statistical models. We conclude with some general remarks on the potential of imprecise probabilities to cope with the multidimensional nature of uncertainty.

Key words and phrases: Imprecise probabilities, Choquet capacities, updating, neighborhood models, generalized Bayes rule, Dempster's rule of conditioning.

1. INTRODUCTION

In their stimulating paper "Judicious Judgment Meets Unsettling Updating: Dilation, Sure Loss and Simpson's Paradox," Ruobin Gong and Xiao-Li Meng (hereafter GM) offer a fresh perspective on famous problems that have long shaken the foundations of statistical analysis. GM manage to trace the paradoxes back to seemingly self-contradictory model assumptions about the marginals and the joint distribution and creatively relate them to phenomena occurring in updating imprecise probabilities. These insights are an excellent example of how the general framework of imprecise probabilities, through its expanded understanding of uncertainty, not only provides new opportunities for statistical modeling, but also helps to illuminate hidden implicit assumptions in classical modeling.

In this short note, we provide in Section 2 some variations of the central topic of conditioning under a generalized probabilistic setting. We will make explicit some mathematical properties of Choquet capacities of order 2 that are contained implicitly in GM's paper. In particular, these properties will allow us to characterize the three different ways of conditioning as envelopes of certain sets of conditional probabilities. In the light of this characterization, we will revisit the notions of "being cautious" and "overfitting," contrasting the generalized Bayes rule (GBR) and Dempster's rule as extreme positions that also allow generalizations by taking an intermediate position. In Section 3, we will address the question of how general the assumed model class of Choquet capacities of order 2 actually is, and thus which practically relevant models are covered by it. Section 4 is reserved for some general concluding remarks on the potential of imprecise probabilities in the context of complex uncertainty.

Thomas Augustin is Professor of Statistics and Head of the Foundations of Statistics and their Applications Group, Department of Statistics, Ludwig-Maximilians Universität München (LMU Munich), Ludwigstr. 33, D-80539 Munich, Germany (e-mail: thomas.augustin@stat.uni-muenchen.de). Georg Schollmeyer is a post-doctorial staff member of the Foundations of Statistics and their Applications Group, Department of Statistics, Ludwig-Maximilians Universität München (LMU Munich), Ludwigstr. 33, D-80539 Munich, Germany (e-mail: georg.schollmeyer@stat.uni-muenchen.de).

2. ENVELOPE REPRESENTATIONS OF THE DIFFERENT CONCEPTS OF CONDITIONAL PROBABILITIES

2.1 A Common Representation

Our argumentation below strongly relies on the following lemma, guaranteeing that for Choquet capacities of order 2 the lower and respectively upper probabilities $\underline{P}$ and $\overline{P}$ of chains of events are *simultaneously* attained by a classical probability in the induced set of compatible distributions. (For further reference and in accordance with the literature, we use the term *credal set* (induced by $\underline{P}$ and $\overline{P}$) for this set of compatible distributions in the set $\mathcal{M}$ of all distributions on the considered measurable space.)

LEMMA 1. *Let $\underline{P}$ be a lower probability such that its credal set $\mathcal{P} = \{P \in \mathcal{M}, P \geq \underline{P}\}$ is relatively compact. Then $\underline{P}$ is two-monotone if and only if for every chain of events $E_1 \subseteq E_2 \subseteq E_3 \ldots \subseteq E_n$ there exists a probability $P \in \mathcal{P}$ such that $P(E_i) = \underline{P}(E_i)$ for all $i \in \{1, \ldots, n\}$.*[1]

This lemma is used in GM's paper implicitly, for instance, in the closed-form reformulation of the generalized Bayes rule in (GM, 2.11f) valid for Choquet capacities of order 2. Using it explicitly, and applying it to the events $E_1 = A \cap B$ and $E_2 = B$, shows that the ratios in (GM, 2.8), and in (GM, 2.9), respectively, are simultaneously optimized. Assuming $\underline{P}(B) > 0$ to make all expressions well-defined, this allows us to rewrite the considered types of conditional probabilities in a unified way (cf., e.g., Gilboa and Schmeidler, 1993). We obtain

$$(1)\qquad \begin{aligned} \underline{P}_{\mathfrak{Z}}(A|B) &= \inf_{P \in \mathcal{P}_{\mathfrak{Z}}} \frac{P(A \cap B)}{P(B)} \quad \text{and} \\ \overline{P}_{\mathfrak{Z}}(A|B) &= \sup_{P \in \mathcal{P}_{\mathfrak{Z}}} \frac{P(A \cap B)}{P(B)}, \end{aligned}$$

where

$$(2)\qquad \mathcal{P}_{\mathfrak{Z}} = \begin{cases} \mathcal{P}, & \text{Gen. Bayes rule,} \\ \mathcal{P}_{\mathfrak{D}}, & \text{Dempster's rule,} \\ \mathcal{P}_{\mathfrak{G}}, & \text{Geometric rule,} \end{cases}$$

with

$$\begin{aligned} \mathcal{P} &\overset{\text{def}}{=} \{P \in \mathcal{M} | P \geq \underline{P}\}, \\ \mathcal{P}_{\mathfrak{D}} &\overset{\text{def}}{=} \{P \in \mathcal{M} | [P \geq \underline{P}] \wedge [P(B) = \overline{P}(B)]\}, \\ \mathcal{P}_{\mathfrak{G}} &\overset{\text{def}}{=} \{P \in \mathcal{M} | [P \geq \underline{P}] \wedge [P(B) = \underline{P}(B)]\}. \end{aligned}$$

2.2 Focusing Versus (Strong) Belief Revision

The envelope representation illustrates GM's important distinction between two different conceptualizations of updating, namely updating as belief revision versus updating as focusing (cf., Dubois and Prade, 1997). In focusing, generic knowledge is not changed, instead, it is only applied to the event that corresponds to the observed data. This leads to the generalized Bayes rule. In contrast, in belief revision one modifies generic knowledge or factual evidence about a problem in the light of new knowledge or evidence. Equation (2) underlines that both the geometric rule as well as Dempster's rule perform a rather strong revision, which may also be interpreted as a strong "overfitting." Constructing $\mathcal{P}_{\mathfrak{D}}$ and $\mathcal{P}_{\mathfrak{G}}$, they both rely exclusively on a single value taken from the interval $[\underline{P}(B), \overline{P}(B)]$. While the geometric rule confines itself on the lowest value, Dempster's rule concentrates on the highest one.[2] In a classical parametric Bayesian setting, where the prior distribution of a parameter ϑ is updated, based on sample B, to the corresponding posterior distribution, $P(B)$ is the predictive distribution of the sample. Then Dempster's rule refines the underlying credal set $\mathcal{P}$ to contain only those probabilities giving the sample the highest likelihood. Indeed, Gilboa and Schmeidler (Gilboa and Schmeidler, 1994, Gilboa and Schmeidler, 1993; see also, Dubois and Prade, 1997) denote Dempster's rule as "maximum likelihood update." Moreover, in particular if we understand $\mathcal{P}$ as parameterized by a nuisance parameter, Dempster's rule can be interpreted as an empirical Bayes approach. It corresponds to the so-called *ML-II approach* (e.g., Berger, 1985, Section 3.5.4), originally suggested by Good (see Good, 1983, e.g., p. 46f).

In this sense, one can also conceptually differentiate the generalized Bayes rule and Dempster's rule as an ideal type dichotomy between an optimistic view and a pessimistic/conservative view. While according to the generalized Bayes rule the conditional lower probability is obtained as the worst conditional classical probability that is consistent with the given lower and upper probabilities,

[1] For a proof, see, for example, Chateauneuf and Jaffray (1989), Proposition 12, p. 277.

[2] This argumentation understands, in accordance with GM's paper, Dempster's rules of conditioning and combination as producing a nonadditive set-function enveloping a set of probabilities. To avoid misunderstandings, it may be noted explicitly that in the so-called *Dempster–Shafer Theory of Belief Functions* popular in artificial intelligence this interpretation is strongly rejected by many authors: "Most important, a probability-bound interpretation is incompatible with Dempster's rule for combining belief functions. If we make up numbers by thinking of them as lower bounds on true probabilities, and we then combine these numbers by Dempster's rule, we are likely to obtain erroneous and misleading results." Shafer (1990), p. 335. Then, belief functions derived from Dempster's rule of conditioning, and more generally from Dempster's rule of combination, are understood as providing an uncertainty calculus of its own. (For a recent review, see Denœux, 2016.)

Dempster's rule can be viewed as a very optimistic approach, radically excluding all probability functions that are not maximally plausible in the light of the observed event B. Somewhere within (and to some extent also somewhere beside?) this ideal type dichotomy, the geometric rule can be located as a rule, which, in contrast to the GBR, restricts $\mathcal{P}$ for updating, however, in contrast to Dempster's rule, in a pessimistic way. It restricts $\mathcal{P}$ to all compatible probabilities that assign the lowest possible probability to the observed event B. Although somehow parallel in construction to Dempster's rule, this way of restricting $\mathcal{P}$ by relying on the lowest possible likelihood is a minimax perspective, that is very cautious from the learning point of view. Indeed, quite naturally, this rule, cannot sharpen vacuous prior information (compare Section 4.3 in GM's paper). However, note that the geometric rule does still always gives bounds that are equally sharp as or sharper then the bound of the GBR, because the infimum in (1) is taken over a smaller set.

2.3 Soft Revision and Likelihood Cuts

These deliberations suggest a quite natural compromise between optimism and pessimism, between the conservative focusing on one hand and the strong revision of Dempster's rule (and the geometric rule) on the other hand, which can be suspected to posses a strong tendency towards overfitting. Instead of basing the revision on one of the interval limits of $[\underline{P}(B), \overline{P}(B)]$, one relies on a subinterval of high or small values. More concretely, for a fixed real value $\alpha \in [0, 1]$, one replaces in (2) the condition $P(B) = \overline{P}(B)$ by the condition[3]

$$P(B) \geq \alpha \cdot \overline{P}(B), \tag{3}$$

or dually, the condition $P(B) = \underline{P}(B)$, which is equivalent to $P(B^c) = \overline{P}(B^c)$), by

$$P\big(B^c\big) \geq \alpha \cdot \overline{P}\big(B^c\big) \tag{4}$$

to obtain suitable generalizations of Dempster's rule and the geometric rule, respectively.[4] For $\alpha = 0$, we obtain GBR, and for $\alpha = 1$ we reproduce Dempster's rule or the geometric rule, respectively. In this sense, α can be seen here as a "parameter of revision." For a small, but positive value α, these revision rules do not rigidly revise the model to only the compatible probabilities that give the observed event B the most/least probability. Such soft revisioning rules may be quite attractive when one feels uncomfortable with the overfitting character of strong revision rules.

Soft revision rules are not coherent in the sense of Walley's (1991) general coherence theory justifying the GBR. In fact, the GBR does not perform any revisioning at all; it never changes the priori assessments, but merely focuses on the implication for situations in which B is observed. As discussed in Section 4.3 of GM's paper, one can thus not learn with the GBR from vacuous prior knowledge.[5] In contrast, the α-cut rule with $\alpha > 0$ and congenial rules are able to learn from vacuous priors.

3. ON THE ROLE OF TWO-MONOTONE CHOQUET-CAPACITIES IN STATISTICS

Many of the results in GM's paper build on the condition that the lower and upper probabilities are Choquet capacities of order 2. In this section, we look at the natural question arising how restrictive this assumption is from a statistical modeling perspective.

From the principled standpoint of the general theory of imprecise probabilities, the condition of two-monotonicity seems artificial. Neither in the behavioral approach to imprecise probabilities (see, in particular, Walley, 1991) nor within its frequentist counterpart (developed by Fine and students, e.g., Fierens, Rêgo and Fine, 2009), two-monotonicity has a contextual meaning or natural interpretation. In addition, two-monotonicity plays also no prominent role in the interpretation-independent branch of imprecise probabilities following Weichselberger (2001). Nevertheless, two-monotone lower probabilities are quite attractive for statistics. In particular, following the prominent Huber–Strassen theorem (Huber and Strassen, 1973; see also Augustin, Walter and Coolen, 2014, Section 7.5.2, for a review of work building on it), two-monotone lower probabilities allow for a rigorous generalization of Neyman–Pearson tests to imprecise probabilities.

A very rich class of two-monotone lower probabilities, which historically also motivated the development of the Huber–Strassen theorem, is provided by certain neighborhood models (see, e.g., Augustin and Hable, 2010, Montes, Miranda and Destercke, 2020a, 2020b). They allow, quite attractively, to formalize the notion of "approximately true distributions," for instance, by considering all distributions close to a certain *central distribution* p^*.

[3] This approach has already been introduced by Cattaneo (2014).

[4] Another variant of generalization would be to replace $P(B) = \overline{P}(B)$ by $P(B) \geq \underline{P}(B) + \alpha \cdot (\overline{P}(B) - \underline{P}(B))$ and to replace $P(B) = \underline{P}(B)$ analogously. Other generalizations are, of course, thinkable as well, for instance neighborhood-models around the maximizing/minimizing probabilities. As a further alternative, Held, Augustin and Kriegler (2008) consider a mixture of the layers produced by the different values of $[\underline{P}(A), \overline{P}(A)]$.

[5] To guarantee that GBR-like inferences with vacuous priors lead to nonvacuous posteriors, extreme prior probabilities have to be excluded. This is achieved by the rather prominent *Imprecise Dirichlet Model* (Walley, 1996) for inference from categorical data. Generally, so-called *near-ignorance prior models* can be considered (see, in particular, Benavoli and Zaffalon's, 2015 approach for multivariate exponential families).

Therefore, neighborhood models have been used in particular in robust statistics as an imprecise sampling distribution or in robust Bayesianism as generalized prior distributions. Typical examples include the δ-total variation model, comprising all distributions where the total variation distance to p^* is smaller than δ, or the ϵ-contamination model formalizing the situation where at least $(1-\epsilon)\cdot 100\%$ of the observations follow the central distribution p^*, but the remaining $\epsilon \cdot 100\%$ may just follow any arbitrary distribution. Generally, many neighborhood models can be written in the form $f \circ p^*$, where convexity of f guarantees two-monotonicity.[6]

Other natural ways of constructing two-monotone models are discrete models with bounds on the probabilities of singletons only (*probability intervals*, Weichselberger and Pöhlmann, 1990) or bounds on distribution functions. The latter, often called *p-boxes*, play a prominent role in generalized uncertainty quantification in reliability analysis (see, e.g., Destercke, Dubois and Chojnacki, 2008).

Finally, also a natural connection between the granularity of observation and two-monotonicity shall be mentioned. Given two measurable spaces $(\Omega, \mathcal{A})$ and $(\Omega, \mathcal{F})$ with $\mathcal{F} \supseteq \mathcal{A}$, a two-monotone lower probability $\underline{P}^*$ can be constructed by extending a two-monotone lower probability $\underline{P}$ on $\mathcal{A}$ to events in $\mathcal{F}$ by natural extension (cf. Walley, 1981, p. 52), and the different concepts of conditioning can be applied. Naturally, if $\underline{P}$ is a classical probability and conditioning is performed by considering only partitions in $\mathcal{A}$, all considered concepts of conditioning coincide in this case.

4. SOME GENERAL CONCLUDING REMARKS

From a principled and general perspective, we unanimously share GM's enthusiasm for a generalized understanding and modeling of uncertainty. What had become obvious in the first AI summer in the context of expert systems and general systems theory, is currently even more important in the environment of ubiquitous and widely available data."Uncertainty is a multidimensional concept. [However, its …] multidimensional nature was obscured when uncertainty was conceived solely in terms of [classical] probability theory, in which it is manifested by only one of its dimensions." (Klir and Wierman, 1999, p. 1)

Indeed, as statisticians and data scientists, we have to pay more attention to the so-to-say "big data uncertainty," that is, those dimensions of uncertainty that go beyond sampling uncertainty and thus do not vanish with increasing sample size. Only generalized probabilistic approaches used in a sophisticated way, as in GM's paper, allow to distinguish between variability and indeterminacy, which is crucial for an appropriate modeling of the quality of probabilistic information. These models are naturally imprecise, or—to avoid the unfortunate misnomer "imprecise" for actually better and more accurate models—rather, set-valued. This set-valued character promises to express scarce, conflicting or simply incomplete information without having to rely on unwarranted assumptions. We agree with GM that making strong but untestable assumptions about unobservable structures just for the sake of a seemingly precise result undermines practical relevance of the statistical analysis, well aware of the "Law of Decreasing Credibility,"

> "The credibility of inferences decreases with the strength of the assumptions maintained" (Manski, 2003, p. 1),

as Manski and his followers put it in the area of partial identification, a rather parallel running development of powerful set-valued analysis in econometrics.[7]

REFERENCES

AUGUSTIN, T. and HABLE, R. (2010). On the impact of robust statistics on imprecise probability models: A review. *Struct. Saf.* **32** 358–365.

AUGUSTIN, T., WALTER, G. and COOLEN, F. P. A. (2014). Statistical inference. In *Introduction to Imprecise Probabilities*. *Wiley Ser. Probab. Stat.* 135–189. Wiley, Chichester. MR3287404 https://doi.org/10.1002/9781118763117.ch7

BENAVOLI, A. and ZAFFALON, M. (2015). Prior near ignorance for inferences in the k-parameter exponential family. *Statistics* **49** 1104–1140. MR3378034 https://doi.org/10.1080/02331888.2014.960869

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR0804611 https://doi.org/10.1007/978-1-4757-4286-2

CATTANEO, M. E. G. V. (2014). A continuous updating rule for imprecise probabilities. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*. *Part III*. *Commun. Comput. Inf. Sci.* **444** 426–435. Springer, Cham. MR3616468

CHATEAUNEUF, A. and JAFFRAY, J.-Y. (1989). Some characterizations of lower probabilities and other monotone capacities through the use of Möbius inversion. *Math. Social Sci.* **17** 263–283. MR1006179 https://doi.org/10.1016/0165-4896(89)90056-5

DENŒUX, T. (2016). 40 years of Dempster–Shafer theory [editorial]. *Internat. J. Approx. Reason.* **79** 1–6. MR3548391 https://doi.org/10.1016/j.ijar.2016.07.010

DESTERCKE, S., DUBOIS, D. and CHOJNACKI, E. (2008). Unifying practical uncertainty representations. I. Generalized p-boxes. *Internat. J. Approx. Reason.* **49** 649–663. MR2475260 https://doi.org/10.1016/j.ijar.2008.07.003

[6] Such models are also used in insurance mathematics as *distorted probabilities*; see, for instance, Wang and Young (1998) for premium calculation in this context, where also the GBR and Dempster's rule are discussed.

[7] See also the surveys by Molinari (2020) on the development of partial identification in microeconometrics and Molchanov and Molinari (2018) on the use of random sets in the context of partial identification.

DUBOIS, D. and PRADE, H. (1997). Focusing vs. belief revision: A fundamental distinction when dealing with generic knowledge. In *Qualitative and Quantitative Practical Reasoning* (D. M. Gabbay, R. Kruse, A. Nonnengart and H. J. Ohlbach, eds.) 96–107. Springer, Berlin, Heidelberg.

FIERENS, P. I., RÊGO, L. C. and FINE, T. L. (2009). A frequentist understanding of sets of measures. *J. Statist. Plann. Inference* **139** 1879–1892. MR2497546 https://doi.org/10.1016/j.jspi.2008.08.025

GILBOA, I. and SCHMEIDLER, D. (1993). Updating ambiguous beliefs. *J. Econom. Theory* **59** 33–49. MR1211549 https://doi.org/10.1006/jeth.1993.1003

GILBOA, I. and SCHMEIDLER, D. (1994). Additive representations of non-additive measures and the Choquet integral. *Ann. Oper. Res.* **52** 43–65. MR1293559 https://doi.org/10.1007/BF02032160

GOOD, I. J. (1983). *Good Thinking*: *The Foundations of Probability and Its Applications*. Univ. Minnesota Press, Minneapolis, MN. MR0723501

HELD, H., AUGUSTIN, T. and KRIEGLER, E. (2008). Bayesian learning for a class of priors with prescribed marginals. *Internat. J. Approx. Reason.* **49** 212–233. MR2454840 https://doi.org/10.1016/j.ijar.2008.03.018

HUBER, P. J. and STRASSEN, V. (1973). Minimax tests and the Neyman–Pearson lemma for capacities. *Ann. Statist.* **1** 251–263. MR0356306

KLIR, G. J. and WIERMAN, M. J. (1999). *Uncertainty-Based Information. Elements of Generalized Information Theory*, 2nd ed. *Studies in Fuzziness and Soft Computing* **15**. Physica-Verlag, Heidelberg. MR1719879 https://doi.org/10.1007/978-3-7908-1869-7

MANSKI, C. F. (2003). *Partial Identification of Probability Distributions*. *Springer Series in Statistics*. Springer, New York. MR2151380

MOLCHANOV, I. and MOLINARI, F. (2018). *Random Sets in Econometrics*. *Econometric Society Monographs* **60**. Cambridge Univ. Press, Cambridge. MR3753715 https://doi.org/10.1017/9781316392973

MOLINARI, F. (2020). Microeconometrics with partial identification. In *Handbook of Econometrics*. *Volume* 7, *Part A* (S. N. Durlauf, L. P. Hansen, J. J. Heckman and R. L. Matzkin, eds.) 355–486.

MONTES, I., MIRANDA, E. and DESTERCKE, S. (2020a). Unifying neighbourhood and distortion models: Part I—new results on old models. *Int. J. Gen. Syst.* **49** 602–635. MR4151121 https://doi.org/10.1080/03081079.2020.1778682

MONTES, I., MIRANDA, E. and DESTERCKE, S. (2020b). Unifying neighbourhood and distortion models: Part II—new models and synthesis. *Int. J. Gen. Syst.* **49** 636–674. MR4151122 https://doi.org/10.1080/03081079.2020.1778683

SHAFER, G. (1990). Perspectives on the theory and practice of belief functions. *Internat. J. Approx. Reason.* **4** 323–362. MR1078973 https://doi.org/10.1016/0888-613X(90)90012-Q

WALLEY, P. (1981). Coherent lower (and upper) probabilities Technical Report Univ. Warwick Coventry. Statistics Research Report.

WALLEY, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. *Monographs on Statistics and Applied Probability* **42**. CRC Press, London. MR1145491 https://doi.org/10.1007/978-1-4899-3472-7

WALLEY, P. (1996). Inferences from multinomial data: Learning about a bag of marbles. *J. Roy. Statist. Soc. Ser. B* **58** 3–57. MR1379233

WANG, S. S. and YOUNG, V. R. (1998). Risk-adjusted credibility premiums using distorted probabilities. *Scand. Actuar. J.* **2** 143–165. MR1659301 https://doi.org/10.1080/03461238.1998.10413999

WEICHSELBERGER, K. (2001). *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I*: *Intervallwahrscheinlichkeit als umfassendes Konzept* [*Elementary Foundations of a More General Calculus of Probability I*: *Interval Probability as a Comprehensive Concept*; *in German*]. Physica, Heidelberg.

WEICHSELBERGER, K. and PÖHLMANN, S. (1990). *A Methodology for Uncertainty in Knowledge-Based Systems*. *Lecture Notes in Computer Science* **419**. Springer, Berlin. MR1048906