Comment on Gong and Meng's *Judicious judgment meets unsettling updating: Dilation, sure loss, and Simpson's paradox*[*]

Chuanhai Liu[†] and Ryan Martin[‡]

November 17, 2020

1 Introduction

Ruobin Gong and Xiao-Li Meng are to be congratulated for their thought-provoking article shedding light on the paradoxical results that can surface when imprecise or incompletely-specified models are updated, in light of observed data, using formal rules like Dempster's and generalized Bayes. With scientific problems becoming increasingly more complex, the idea that models describing the phenomena under investigation can be precisely specified is a fantasy, so Gong and Meng's insights about the effects of these updating rules are both important and timely. However, after highlighting a number of cases where the updates are "unsettling," they give no recommendations about which updating rule, if any, is reliable. In some cases, generalized Bayes seems to be the right choice, while in others it's Dempster's rule. Since we can't rely on any of the updating rules to give satisfactory answers in every problem, apparently our only recourse is to use "judicious judgment" on a case-by-case basis.

Here, however, we argue that steps toward settling what's unsettling about these updates can be made by taking a different perspective on what a solution to the statistical problem entails. Gong and Meng make their perspective very clear:

> *Statistical learning is a process through which models perform updates in light of new information, according to a pre-specified set of operation rules.*

What's missing from this description is that inferences drawn based on the updated models must be reliable or valid in some specific sense, otherwise, the results are not useful. So the question is not really about updating beliefs but, rather, how to ensure that the beliefs data scientists construct for inference and prediction achieve the desired reliability properties. From this perspective, Gong and Meng's goal is overly ambitious: for valid and efficient inference, rules that update beliefs are not necessary. A less ambitious goal—but still in line with the priorities of scientists—is to understand what it takes to

[*]Prepared for inclusion in the discussion scheduled to appear in *Statistical Science*.

[†]Department of Statistics, Purdue University, chuanhai@purdue.edu

[‡]Department of Statistics, North Carolina State University, rgmarti3@ncsu.edu

construct procedures for allocating beliefs such that inferences drawn are *valid* and *efficient*. The first step is to define what these terms mean, which we do below in Section 2. We immediately take comfort in the fact that validity rules out the troubling sure loss phenomenon, and, as we show in Section 3, validity and efficiency make it possible to compare the solutions based on the different updating rules. Of course, if validity and efficiency are the goal, then it makes sense to follow a procedure that is specifically design to achieve these properties. The *inferential model* (IM) procedure introduced in Martin and Liu (2013, 2015b) is just that, and in Section 4 we describe this framework and show how it generally leads to better solutions than those based on the formal updating rules in Gong and Meng's examples. The take-away message is that, by following the validity- and efficiency-focused IM approach, the "unsettling" phenomena identified by Gong and Meng can be avoided. Finally, Section 5 gives some concluding remarks and mentions a few topics for future investigation.

2 Valid and efficient prediction

The examples in Gong and Meng (2020) are most conveniently described as prediction problems, so that's the perspective we take; all of this can be developed in a similar way for statistical inference. To set the scene, let X denote the observable data and $Y \in \mathbb{Y}$ the quantity to be predicted. Next, let P denote the probability measure that describes the joint distribution of (X, Y), at least partially unknown or unspecified. As indicated above, we proceed by quantifying uncertainty about Y, given $X = x$, via a pair of lower and upper probabilities, denoted by $(\underline{\pi}_x, \overline{\pi}_x)$, defined on $\mathbb{Y}$. We refer to the map $x \mapsto (\underline{\pi}_x, \overline{\pi}_x)$ as a *probabilistic predictor*, and the user's degree of belief in the truthfulness of an assertion $A \subseteq \mathbb{Y}$ concerning the unobserved Y, given $X = x$, are described by the pair $(\underline{\pi}_x(A), \overline{\pi}_x(A))$. Note that the probabilistic predictor need not be based on updating a precise or imprecise probability model.

Since the goal is for the probabilistic predictor to make reliable predictions, i.e., not wrong too often, consider the following prediction validity property.

Definition (Cella and Martin 2020). A probabilistic predictor is *valid* if

$$\mathsf{P}\{\overline{\pi}_X(A) \leq \alpha \,,\, Y \in A\} \leq \alpha, \quad \forall\, (A, \alpha, \mathsf{P}), \tag{1}$$

where the probability is with respect to the joint distribution of (X, Y) determined by P and "$\forall$" is over all assertions $A \subseteq \mathbb{Y}$, all levels $\alpha \in [0, 1]$, and all joint distributions P.

The intuition is that, at least for small α, the data analyst interprets the event "$\overline{\pi}_X(A) \leq \alpha$" as evidence against the truthfulness of the assertion A about Y, so the joint event "$\overline{\pi}_X(A) \leq \alpha, Y \in A$" is one where an erroneous prediction is possible. Then (1) requires that the user be able to control the frequency of such erroneous predictions. Thanks to the familiar duality between lower and upper probabilities, a similar condition can be formulated in terms of $\underline{\pi}_x$ (Cella and Martin 2020). To see what condition (1) imposes on the probabilistic predictor, consider the equivalent expression

$$\mathsf{E}\{1_{\overline{\pi}_X(A) \leq \alpha}\, \mathsf{P}(Y \in A \mid X)\} \leq \alpha, \quad \forall\, (A, \alpha, \mathsf{P}), \tag{2}$$

where 1_B is the indicator function, E is expectation with respect to the marginal distribution of X under P, and $\mathsf{P}(Y \in A \mid X)$ is the conditional probability based on P. Clearly,

if $\overline{\pi}_x(A)$ equals or dominates the conditional probability $\mathsf{P}(Y \in A \mid x)$ or the marginal probability $\mathsf{P}(Y \in A)$, then (2) holds. This connection between validity and "dominance" leads to several interesting observations, as discussed in Cella and Martin (2020).

- Sure loss, the most unsettling of the three phenomena studied by Gong and Meng, is ruled out by validity, that is, validity implies no sure loss.
- If the imprecise model is known to contain the true joint distribution of (X, Y), like in Gong and Meng's examples, then the generalized Bayes solution is valid.

While generalized Bayes provides a strategy to achieve validity, it's not the only available strategy and often will not be the best. We'll have more to say about this below.

Beyond validity, efficiency is important too. Here, we say that between a pair of valid probabilistic predictors, with upper probabilities $\overline{\pi}_x$ and $\overline{\pi}'_x$, the latter is no less efficient than the former—with respect to a specified assertion A—if $\overline{\pi}'_x(A) \leq \overline{\pi}_x(A)$ for all x. The idea is that large upper probabilities are trivially valid, so the goal is to find the smallest possible upper probabilities that satisfy (1) or (2). By the duality between lower and upper probabilities, similar intuition can be developed for $\underline{\pi}_x$. We'll not investigate validity or efficiency formally here, only in the context of two examples in Section 3.

3 Gong and Meng's examples

3.1 Three prisoners

Three prisoners—labeled A, B, and C—are in custody and one will be randomly chosen to have their sentence pardoned; the other two will be executed. Let Y denote the pardoned prisoner. Prisoner A ask the guard to tell him which of Prisoners B or C will be executed, and the guard's response is the data X. The goal is to predict Y based on data X. What do the validity and efficiency considerations add to the discussion?

As Gong and Meng argue, the joint distribution for (X, Y) is fully determined except for the conditional probability $\theta = \mathsf{P}(X = \text{B} \mid Y = \text{A})$. So, for the most relevant assertion, "$Y = \text{A}$," the validity condition (2) can be expressed as

$$1_{\overline{\pi}_{\text{B}}(\text{A}) \leq \alpha} \cdot \tfrac{\theta}{3} + 1_{\overline{\pi}_{\text{C}}(\text{A}) \leq \alpha} \cdot \tfrac{1-\theta}{3} \leq \alpha. \tag{3}$$

As presented in Gong and Meng—see, also, Walley (1991, Sec. 6.4.4)—the generalized Bayes solution returns a probabilistic predictor with

$$\underline{\pi}_x(\text{A}) = 0 \quad \text{and} \quad \overline{\pi}_x(\text{A}) = \tfrac{1}{2}, \quad x \in \{\text{B}, \text{C}\},$$

and, for this, it's easy to check that (3) holds. Dempster's rule returns a probabilistic predictor with *lower and upper* probabilities for "$Y = \text{A}$" equal to $\frac{1}{2}$, for all x. This satisfies (3) at "$Y = \text{A}$," but not if we consider the complementary assertion. Indeed, with Dempster's probabilistic predictor at the assertion "$Y \in \{\text{B},\text{C}\}$,", the validity requirement in (3) boils down to

$$1_{\frac{1}{2} \leq \alpha} \cdot \tfrac{2}{3} \leq \alpha. \tag{4}$$

Taking $\alpha = \frac{1}{2}$ leads to a contradiction. This is basically the proof of how sure loss leads to a violation of validity in general. Similarly, the solution based on the geometric rule, which also suffers from sure loss in this example, is invalid.

A closer look at (3) provides some insight as to what the "most efficient" solution is. If $\underline{\pi}_x(\text{A}) = \frac{1}{3}$ for each $x \in \{\text{B}, \text{C}\}$, then (3) would be satisfied, and it would be more efficient than the generalized Bayes solution. It would also be valid since lower probability on the complementary event is $\frac{2}{3}$, as opposed to Dempster's $\frac{1}{2}$, so it would not get caught by the trap (4). We'll see below how this "most efficient" solution can be achieved.

3.2 Boxer, wrestler, and coin

Let Y_1 denote the outcome a fair coin flip, with $Y_1 = 1$ and $Y_1 = 0$ corresponding to Heads and Tails, respectively, and let Y_2 denote the outcome of the boxer versus wrestler match, with $Y_1 = 1$ and $Y_1 = 0$ denoting a boxer and wrestler victory, respectively. The data is $X = |Y_1 - Y_2|$, an indicator that Y_1 and Y_2 take the same value. The goal is to predict the outcome of the fight (or of the coin flip) based on the observed value of X.

Features of the joint distribution of (X, Y), with $Y = (Y_1, Y_2)$, are left unspecified, in particular, the conditional probabilities

$$\theta_{1|y_1} = \mathsf{P}(Y_2 = 1 \mid Y_1 = y_1), \quad y_1 \in \{0, 1\}.$$

This pair $\theta = (\theta_{1|0}, \theta_{1|1})$ of conditional probabilities can take any value in $[0,1]^2$. That is, the problem setup doesn't rule out the possibility that the fight's outcome is determined by the coin flip, or that the fight's outcome is independent of the coin and pre-determined.

As above, let's start by specializing the validity condition to the present example. That is, if $\overline{\pi}_x(1)$ is the probabilistic predictor's upper probability at the assertion "$Y_2 = 1$," i.e., a boxer victory, then (2) requires

$$\tfrac{1}{2}\big\{1_{\overline{\pi}_0(1) \leq \alpha} \cdot \theta_{1|0} + 1_{\overline{\pi}_1(1) \leq \alpha} \cdot \theta_{1|1}\big\} \leq \alpha.$$

Since $(\theta_{1|0}, \theta_{1|1})$ can take any value in $[0,1]^2$, there is no way to ensure that validity holds, except trivially, by taking the upper probabilities identically equal to 1. This is precisely the generalized Bayes solution in Gong and Meng. Dempster's rule, again, is invalid.

For assertions about the coin, surprisingly, the only satisfactory solution based on the methods investigated in Gong and Meng is that based on Dempster's rule, which ignores the data and uses the known marginal distribution of Y_1. It's easy to check that the simple probabilistic predictor

$$\underline{\pi}_x(\text{"}Y_1 = 1\text{"}) = \overline{\pi}_x(\text{"}Y_1 = 1\text{"}) = \tfrac{1}{2}, \quad x \in \{0, 1\},$$

is valid and efficient. We'll see below how this solution can be achieved in the IM context.

4 Inferential models

4.1 Formulation

The IM formulation starts by specifying an *association* between what is being modeled, i.e., data X and quantity of interest Y, the unknown parameter $\theta \in \Theta$, and an unobservable auxiliary variable U, whose distribution P_U is known, via an equation or rule

$$(X, Y) = a(\theta, U), \quad U \sim \mathsf{P}_U. \tag{5}$$

The mapping $a(\theta, \cdot)$ implicitly encodes what is known about the joint distribution but explicitly depends on the unknown θ. The details depend on the objectives of the analysis: if (X, Y) is observable and the goal is inference on θ, then we proceed as described in Martin and Liu (2013, 2015b); if only X is observable and the goal is prediction of Y, then we proceed as in Martin and Lingham (2016) or Cella and Martin (2020).

For the case of prediction, the idea is as follows. Given $X = x$, define a set-valued mapping $u \mapsto Q_x(u)$, into the space $\mathbb{Y} \times \Theta$ of unknown quantities, as

$$Q_x(u) = \{(y, \vartheta) \in \mathbb{Y} \times \Theta : (x, y) = a(\vartheta, u)\}.$$

If u satisfies the equation (5) with $X = x$, then $Q_x(u)$ contains the correct prediction. It is impossible to know for sure which u values satisfy the equation, but it is possible—since the distribution P_U is known—to construct a random set $\mathcal{U}$ of u values that we believe is likely to contain a solution. For such a $\mathcal{U}$, the new random set

$$Q_x(\mathcal{U}) = \bigcup_{u \in \mathcal{U}} Q_x(u),$$

obtained by mapping through the association to the space of unknowns, is equally likely to contain the correct prediction. Then we can define the lower and upper probabilistic predictor for Y, given $X = x$,

$$\begin{aligned} \underline{\pi}_x(A) &= \mathsf{P}_{\mathcal{U}}\{Q_x(\mathcal{U}) \subseteq A \times \Theta\} \\ \overline{\pi}_x(A) &= \mathsf{P}_{\mathcal{U}}\{Q_x(\mathcal{U}) \cap (A \times \Theta) \neq \varnothing\}, \end{aligned}$$

where $\mathsf{P}_{\mathcal{U}}$ is the distribution of the random set $\mathcal{U}$ and A is an arbitrary subset of $\mathbb{Y}$. The appropriate choice of random set $\mathcal{U}$ is beyond the scope of this short note, but suffice it to say that choosing $\mathcal{U} \sim \mathsf{P}_{\mathcal{U}}$ to achieve the validity condition is relatively straightforward; see Martin and Liu (2013, 2015b) for details.

The above lower and upper prediction probabilities are belief and plausibility functions, respectively, defined on the power set of $\mathbb{Y}$, determined by the association, data, and user-defined random set. Our focus is on validity and efficiency, so we don't obligate ourselves to manipulating these functions using the Dempster–Shafer calculus of belief functions (Dempster 2008; Shafer 1976). Instead, the focus is on expressing the association between data and unknowns in terms of an auxiliary variable whose dimension is as small as possible. When the dimension is lower, the size of the random set needed to achieve validity is smaller, hence greater efficiency. General strategies for reducing the dimension were presented in Martin and Liu (2015a,c). The marginalization techniques in the latter reference will be used below.

4.2 Three prisoners

For an IM solution, start with an association

$$\begin{aligned} Y &= U_1 \\ X &= f(\theta, U_1, U_2), \end{aligned}$$

where $U_1 \sim \mathsf{Unif}(\{\text{A}, \text{B}, \text{C}\})$ and $U_2 \sim \mathsf{Unif}(0,1)$ are independent, and

$$f(\theta, u_1, u_2) = \begin{cases} \text{B} & \text{if } u_2 \leq 1_{u_1=\text{C}} + \theta \, 1_{u_1=\text{A}} \\ \text{C} & \text{otherwise.} \end{cases}$$

A unique feature of this problem is that the quantity of interest, Y, the identity of the pardoned prisoner, has a known marginal distribution.

Since θ is not of primary interest, there is an opportunity to potentially reduce the auxiliary variable dimension before carrying out the IM construction (Martin and Liu 2015c). Indeed, it is easy to check that, for every (x, y, u_2), there exists a θ such that $x = f(\theta, y, u_2)$. By the general IM marginalization theory, this implies the second equation in the association can be *effectively* ignored. This means valid (and efficient) prediction of Y should proceed based on its known marginal distribution. We say the second equation can be "effectively" ignored because it wouldn't make sense to predict that, say, $Y = \text{B}$ if we observe $X = \text{B}$. So we should account for this information in some way.

Based on the argument above, the A-step concludes by writing $Y = U$, where $U \sim \mathsf{Unif}(\{\text{A}, \text{B}, \text{C}\})$. For the P-step, we introduce a suitable random set $\mathcal{U} \sim \mathsf{P}_{\mathcal{U}}$ targeting the unobserved value of U. There are many options, but here we recommend to take $\mathcal{U}$ with support $\big\{\{\text{B}, \text{C}\}, \{\text{A}, \text{B}, \text{C}\}\big\}$ and masses assigned as

$$\mathsf{P}_{\mathcal{U}}(\mathcal{U} = \{\text{B}, \text{C}\}) = \tfrac{2}{3} \quad \text{and} \quad \mathsf{P}_{\mathcal{U}}(\mathcal{U} = \{\text{A}, \text{B}, \text{C}\}) = \tfrac{1}{3}.$$

With this choice, the probabilistic predictor returned by the IM's C-step is precisely the one described at the end of Section 3.1, the one that is valid and most efficient, superior to all the solutions presented in Gong and Meng (2020) based on updating the imprecise model according to formal rules.

4.3 Boxer, wrestler, and coin

For an IM solution, define an association as

$$Y_1 = 1_{U_1 \leq 0.5} \quad \text{and} \quad Y_2 = 1_{U_2 \leq \theta_{1|1}, U_1 \leq 0.5} + 1_{U_2 \leq \theta_{1|0}, U_1 > 0.5},$$

with $X = |Y_1 - Y_2|$ and (U_1, U_2) a pair of independent $\mathsf{Unif}(0,1)$ random variables. Suppose, for example, that $X = 0$ is observed, i.e., that the outcomes of the fight and coin flip are the same; the case with $X = 1$ is analogous. When X is observed, the outcome of the fight determines the coin flip, and vice versa, so there's no need to consider both Y_1 and Y_2 after X is observed. We start with the case of Y_2, the fight's outcome. A generic (u_1, u_2) is pushed through the assertion, with $X = 0$, to a set in the (y_2, θ)-space:

$$Q_0(u_1, u_2) = \begin{cases} \{(1, \theta) : u_2 \leq \theta_{1|1}\} & \text{if } u_1 \leq 0.5 \\ \{(0, \theta) : u_2 > \theta_{1|0}\} & \text{if } u_1 > 0.5. \end{cases}$$

Since we're only interested in Y_2, our assertions about (Y_2, θ) take the form $\{y_2\} \times [0,1]^2$, for $y_2 \in \{0, 1\}$. We'll leave out the details here, but it can be shown that, for any suitable random set $\mathcal{U} \subseteq [0,1]^2$, the probabilistic predictor for Y_2 returned by the IM is vacuous, i.e., its lower and upper probabilities are 0 and 1, respectively. As we showed above, this is the only valid solution.

Finally, if interest was in predicting Y_1, the outcome of the coin flip, then we could proceed very much like in the three prisoners example. That is, the general theory of marginal inference in Martin and Liu (2015c) allows us to ignore everything except Y_1, hence valid and efficient inference is achieved by using the marginal distribution of Y_1 to construct a valid and efficient probabilistic predictor. This agrees with the solution based on Dempster's rule and is more efficient than that based on the generalized Bayes rule.

5 Conclusion

The examples in Gong and Meng's paper are simultaneously both simple and challenging, making them ideal cases to test our understanding and to highlight the benefits of our perspective that focuses specifically on the construction of data-dependent beliefs that are both valid and efficient. This note is already too long, so we'll present our investigation into Simpson's paradox from an IM perspective elsewhere.

It's interesting to see that, at least in cases where the imprecise model is known to be correctly specified, generalized Bayes is valid. But even in these relatively simple examples, we find that the IM solution can lead to more efficient prediction. In more complex settings, there the generalized Bayes solution faces certain challenges, in particular, specifying an imprecise model that is both sufficiently flexible and simple enough to compute the lower/upper envelopes. So there are ample reasons to consider alternative solutions. For example, Cella and Martin (2020) established a connection between valid IMs and the powerful conformal prediction machinery (Vovk et al. 2005).

Finally, as we were preparing this discussion piece, it occurred to us that the failure of Fisher's fiducial argument and Dempster's extension to achieve valid inference and prediction in general could possibly be understood and expressed in terms of the contraction, dilation, and/or sure loss examined by Gong and Meng. This claim, too, will be investigated further and our results will be presented elsewhere.

References

Cella, L. and Martin, R. (2020). Strong validity, consonance, and a new characterization of conformal prediction. `arXiv:2001.09225`.

Dempster, A. P. (2008). The Dempster–Shafer calculus for statisticians. *Internat. J. Approx. Reason.*, 48(2):365–377.

Gong, R. and Meng, X.-L. (2020). Judicious judgment meets unsettling updating: Dilation, sure loss, and Simpson's paradox. *Statist. Sci.*, to appear, `arXiv:1712.08946`.

Martin, R. and Lingham, R. T. (2016). Prior-free probabilistic prediction of future observations. *Technometrics*, 58(2):225–235.

Martin, R. and Liu, C. (2013). Inferential models: a framework for prior-free posterior probabilistic inference. *J. Amer. Statist. Assoc.*, 108(501):301–313.

Martin, R. and Liu, C. (2015a). Conditional inferential models: combining information for prior-free probabilistic inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 77(1):195–217.

Martin, R. and Liu, C. (2015b). *Inferential Models: Reasoning with Uncertainty*, volume 147 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL.

Martin, R. and Liu, C. (2015c). Marginal inferential models: prior-free probabilistic inference on interest parameters. *J. Amer. Statist. Assoc.*, 110(512):1621–1631.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer, New York.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*, volume 42 of *Monographs on Statistics and Applied Probability*. Chapman & Hall Ltd., London.