Incorporating Expert Opinion in an Inferential Model While Maintaining Validity

Leonardo Cella LOLIVEI@NCSU.EDU
Ryan Martin RGMARTI3@NCSU.EDU
Department of Statistics, North Carolina State University, USA

Abstract

The incorporation of partial prior information in statistical inference problems still lacks a definitive answer. The two most popular statistical schools of thought deal with partial priors in different ways: they either get completely ignored (frequentist approach) or they are transformed into a "complete" prior information, i.e., a probability distribution (Bayesian approach). Acknowledging the importance of (i) taking into account all sources of relevant information in a given problem and (ii) controlling error probabilities, the present paper provides insights on how to incorporate partial priors "as they are". This incorporation is guided by desired properties, such as that correct partial priors should result in more efficient inferences and, most importantly, that the inferences are always calibrated, independent of the truthfulness of the partial prior.

Keywords: Dempster's rule; elastic; plausibility contour; prior distribution; random set.

1. Introduction

The two dominant schools of thought in statistics—frequentist and Bayesian—differ in a number of important ways. One distinguishing feature is how the two handle prior information, expert opinion, etc. Roughly, the former has no formal mechanism for incorporating such information and no need for one since its focus is on error rate control, while the latter explicitly requires such information to be provided in order to carry out the analysis. The reality is that relevant prior information is often available, but it may not be at a sufficient resolution to determine a complete prior distribution and, moreover, it may not be 100% reliable. Then the two classical approaches just described correspond to two extreme ways of dealing with this situation: the frequentists would likely ignore this partial or incomplete prior information, while the Bayesian must fill in the gaps with extraneous details to determine a complete prior probability distribution. Currently, there is no middle-ground, no approach that can (a) incorporate the available prior information as is, without embellishment, and (b) dynamically limit the degree of commitment assigned to the prior information so that desirable calibration properties, i.e., error rate control, can be maintained.

Filling this gap is an ambitious goal indeed, so the present paper has a more modest aim, namely, to provide some important insights as to how this ideal can be accomplished. Let $Y \in \mathbb{Y}$ denote the observable data and $\mathscr{M} = \{\mathsf{P}_{Y|\theta} : \theta \in \Theta\}$ the posited statistical model, which is just a family of probability distributions on $\mathbb{Y}$ indexed by a parameter $\theta \in \Theta$. Then the objective in this statistical inference context is to quantify uncertainty about θ based on an observation $Y = y$ relative to model $\mathscr{M}$. Towards this, we consider an *inferential model*, which is just a map

$$(y, \mathscr{M}, \ldots) \mapsto b_y : 2^{\Theta} \to [0,1],$$

that takes data, model, and perhaps other inputs (e.g., prior information) to a function b_y such that, for any hypothesis $A \subseteq \Theta$ pertaining to the unknown parameter θ, $b_y(A)$ represents the data analyst's degree of belief about the truthfulness of A relative to the observed data y, model $\mathscr{M}$, and other inputs. Familiar things, such as Bayesian posterior distributions, are inferential models, as are some less familiar things like fiducial (e.g., Fisher, 1973; Zabell, 1992), generalized fiducial (e.g., Hannig et al., 2016), and confidence distributions (e.g., Xie and Singh, 2013; Schweder and Hjort, 2016). The aforementioned examples correspond to additive degrees of belief, i.e., probabilities, but inferential models can accommodate non-additive beliefs as well, including belief functions (e.g., Shafer, 1976; Dempster, 2008; Kohlas and Monney, 1995) and possibility measures (e.g., Dubois and Prade, 1988). Special non-additive beliefs that feature frequentist-style calibration properties have attracted some attention in Balch (2012), Martin and Liu (2013, 2016), and Denœux and Li (2018). Recently, Balch et al. (2017) showed that only non-additive beliefs can avoid what they call *false confidence*, i.e., a tendency to assign high degree of belief to certain false hypotheses; see, also, Martin (2019). Our view is that calibration properties are essential to the logic of statistical inference, in order to avoid "systematically misleading conclusions" (Reid and Cox, 2015), so we focus on non-additive inferential models with the desirable *validity* property described in Section 2.

Given that we have already committed to working in the domain of non-additive beliefs, it is very natural that we use this same framework to describe the prior information or, rather, expert opinion. Here we adopt the latter terminology to distinguish our perspective from the Bayesians' where prior information equals a prior distribution, and to emphasize that it really is a belief in both the scientific/epistemological and mathematical senses; unlike in the Bayesian setting where a prior is necessary, here there would be no reason to introduce such a thing if not that an expert assigns a certain degree of belief to it. Finally, as we describe in Section 3, we have two belief functions—one represents expert opinion and the other our inferential model output based on data y, etc.—and the goal is to combine them. For combining belief functions, there are a host of existing strategies, perhaps the most notable being *Dempster's rule of combination* (e.g., Shafer, 1976, Chap. 3). However, the task here is not so straightforward. Remember, the goal is to incorporate expert opinion while preserving the inferential model's desirable validity property. But Dempster's rule has no such considerations built in, so, not surprisingly, it fails to preserve validity. Therefore, new ideas are needed.

The challenge is that expert opinion and the validity property generally are at odds with one another. To make this clear, consider an analogy in the familiar Bayesian context. It is straightforward to construct a confidence interval for a normal mean with exact coverage probability properties but, if a meaningful, proper prior distribution is available and combined with the likelihood via Bayes's theorem, then the corresponding posterior yields a credible interval that is different and generally lacks the desired coverage probability properties. That is, the prior generally has a biasing effect that interferes with the frequentist coverage probability property. This interference is most severe when the prior is "wrong," assigns mass away from the true parameter value. However, if the prior is "right" in the sense that it assigns mass near the true parameter value, then the biasing effect is beneficial and leads to an efficiency gain, i.e., narrower credible intervals that still achieve the desired coverage probability. Of course, we are unable to determine if the prior is "right" or "wrong" and Bayes's theorem does not offer a way to use the data to dynamically weight how much of the prior is used. So here we propose a different type of combination rule that allows for this kind of dynamic weighting.

The specific proposal in this paper is to extend the approach in Ermini Leaf and Liu (2012) that relies on an appropriate stretching of an underlying random set to accommodate a constraint in the parameter space. Here, the expert opinion is more vague than a parameter constraint so more care is needed to determine an appropriate amount of stretching. Moreover, while the validity property is automatically preserved when the random set is stretched, there is a corresponding loss of efficiency. This suggests that there is an opportunity to improve efficiency by performing a second adjustment to the random set, one that contracts in the opposite direction of the stretching. If tuned correctly, this dually elastic random set can lead to inference that incorporates both certain and less-than-certain expert opinion, while maintaining validity and efficiency.

After a review of inferential models and the validity property in Section 2 and a discussion of how the expert opinion can be encoded in terms of a random set in Section 3, we present our specific strategy for stretching and contracting the random set, focused on a specific and simple normal mean problem. Numerical results in Section 5 confirm our claims that validity and efficiency can be achieved for all parameters and for any kind of expert opinion, good or bad. Some concluding remarks are given in Section 6.

2. Valid Inferential Models

As mentioned in Section 1, an inferential model is simply a rule by which the observed data, posited statistical model, etc., can be converted into degrees of belief about the unknown parameter of interest. What makes this construction challenging is that we require our inferential model to have certain calibration properties. Following Martin and Liu (2013), we say that an inferential model is *valid* if its b_y output satisfies

$$\sup_{\theta \notin A} \mathsf{P}_{Y|\theta}\{b_Y(A) \geq 1-\alpha\} \leq \alpha, \qquad \begin{cases} \forall\, \alpha \in (0,1), \\ \forall\, A \subseteq \Theta. \end{cases} \tag{1}$$

That is, false hypotheses tend to be assigned relatively low belief with respect to the posited statistical model. This prevents systematically misleading conclusions.

Define the dual function $p_y(A) = 1 - b_y(A^c)$. We will refer to b_y and p_y as belief and plausibility functions, respectively. Since (1) covers all hypotheses, an equivalent condition can be expressed in terms of plausibility:

$$\sup_{\theta \in A} \mathsf{P}_{Y|\theta}\{p_Y(A) \leq \alpha\} \leq \alpha, \qquad \begin{cases} \forall\, \alpha \in (0,1), \\ \forall\, A \subseteq \Theta. \end{cases} \tag{2}$$

Similarly, this property says that hypotheses which are not false will tend to be assigned relatively high plausibility. An interesting and practically useful consequence is that procedures derived from the plausibility function have frequentist error rate control guarantees. For example, in the simulations presented in Section 5, we will be considering interval estimation and we will make use of the fact the set $\{\vartheta : p_y(\{\vartheta\}) > \alpha\}$ has frequentist coverage probability at least $1-\alpha$ when the validity property holds.

It is not obvious that an inferential model satisfying (1) exists. However, a construction is possible and, to our knowledge, what follows is the only one available. The key idea is to first introduce, in the *association step*, a set of unobservable auxiliary variables, denoted by U, connected to the observable data Y and unknown parameter θ. Next, a random set on the auxiliary variable space is used in the *prediction step* to "guess" the unobserved value of U.

Finally, in the *combination step*, the observed data, the statistical model, and the random set are fused together to create a new random set on Θ, whose distribution is used to determine the data analyst's degrees of belief. These three steps are detailed below.

A-step *Define an* association *consistent with the posited statistical model, i.e., a function a such that data Y from distribution $\mathsf{P}_{Y|\theta}$ can be simulated by the algorithm*

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

where $U \in \mathbb{U}$ is an auxiliary variable and its distribution, P_U, does not depend on any parameters. Given a, define the set-valued maps

$$\Theta_y(u) = \{\vartheta : y = a(\vartheta, u)\}, \quad u \in \mathbb{U}.$$

P-step *Introduce a suitable random set $\mathcal{S}$, with distribution $\mathsf{P}_{\mathcal{S}}$, taking values in $2^{\mathbb{U}}$, designed to* predict *the unobserved value of the auxiliary variable U.*

C-step *Finally,* combine *Θ_y and $\mathcal{S}$ to get a new random set*

$$\Theta_y(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \Theta_y(u). \tag{3}$$

Then the distribution of $\Theta_y(\mathcal{S})$, as a function of $\mathcal{S} \sim \mathsf{P}_{\mathcal{S}}$, for fixed y, determines the inferential model output:

$$b_y(A) = \mathsf{P}_{\mathcal{S}}\{\Theta_y(\mathcal{S}) \subseteq A\}. \tag{4}$$

It turns out that the validity property (1) holds for this belief function output under relatively weak conditions on the random set $\mathcal{S}$. In fact, according to Martin and Liu (2013), a sufficient condition for validity is that $\gamma(U) \sim \mathsf{Unif}(0,1)$, when $U \sim \mathsf{P}_U$, where $\gamma(u) = \mathsf{P}_{\mathcal{S}}(\mathcal{S} \ni u)$ is the plausibility contour of the random set $\mathcal{S}$. This condition links the distribution of $\mathcal{S}$ to the distribution of U, which of course is natural if $\mathcal{S}$ is supposed to be good at predicting realizations from P_U.

As our running example, let Y denote a scalar observable with distribution $\mathsf{N}(\theta, n^{-1})$, i.e., a normal distribution with mean θ and variance n^{-1}; such a situation might arise when an independent and identically distributed sample of size n from $\mathsf{N}(\theta, 1)$ is available and summarized by the sample mean. For the A-step, a natural choice is

$$Y = \theta + U, \quad U \sim \mathsf{P}_U = \mathsf{N}(0, n^{-1}).$$

This association defines a set $\Theta_y(u)$ of candidate parameter values corresponding to the observed y and a generic u, namely,

$$\Theta_y(u) = \{\vartheta : y = a(\vartheta, u)\} = \{y - u\},$$

a singleton set in this case. For the P-step, we introduce a random set $\mathcal{S}$ of the form

$$\mathcal{S} = \{u : |u| \leq |\tilde{U}|\}, \quad \tilde{U} \sim \mathsf{P}_U = \mathsf{N}(0, n^{-1}). \tag{5}$$

This random set satisfies that sufficient condition mentioned above and, in fact, it happens to be "optimal" as in Martin and Liu (2013). Finally, for the C-step, we combine $\Theta_y(\cdot)$ and the random set $\mathcal{S}$ to get

$$\Theta_y(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \Theta_y(u) = (y - |\tilde{U}|, y + |\tilde{U}|), \quad \tilde{U} \sim \mathsf{N}(0, n^{-1}).$$

Write F_n for the $\mathsf{N}(0, n^{-1})$ distribution function, so that the plausibility contour can be written as

$$p_y(\{\vartheta\}) = \mathsf{P}_{\mathcal{S}}\{\Theta_y(\mathcal{S}) \ni \vartheta\} = 1 - |2F_n(y - \vartheta) - 1|.$$

3. Expert Opinions

Validity of the inferential model output is achieved without the assistance of any prior information about the parameter of interest θ. However, as it was discussed in Section 1, the scenario where subject-matter experts have opinions/beliefs about θ is not uncommon. Intuitively, all relevant information should be used as input in an inferential model, and, certainly, an opinion given by an expert is a relevant information. Moreover, if this incorporation is done in a way that the validity property is maintained, i.e., systematic misleading conclusions are avoided even when the expert opinion is inconsistent with the data, then it would be hard to imagine objections to the inclusion of this opinion.

Here we are considering *expert opinion* to be any type of prior information about θ that is not (or cannot be unambiguously translated to) a probability distribution. There is a long list of different ways expert opinions can be formulated, but of particular interest here is the case where they are represented by a (subjective) probability β attached to a fixed, proper subset B of Θ, i.e.,

$$\textit{expert assigns belief probability } \beta \in (0,1] \textit{ to } B. \tag{6}$$

What we have in mind are (realistic) situations in which the expert says "I'm 95% sure that θ is between a and b." The key point is that the expert provides no information about how the probability β is allocated *within* the set B, so no finer-resolution statements about probabilities assigned to subsets of B can be made. The Bayesian, however, must choose a full prior distribution for θ that accommodates the expert opinion, but this requires something extra beyond the information provided by the expert, hence some potential for bias that can negatively affect inference.

The ultimate goal of the present study is to incorporate to the already valid inferential model the expert opinion in (6) *as is*, without embellishment. But the following question arises: how to represent the expert's low-resolution opinion mathematically? It turns out that the theory of random sets (e.g., Molchanov, 2005; Nguyen, 2006) naturally accommodates expert opinions in the form of (6). Consider the random set $\mathscr{E}$, where

$$\mathsf{P}_{\mathscr{E}}(\mathscr{E} = B) = \beta \quad \text{and} \quad \mathsf{P}_{\mathscr{E}}(\mathscr{E} = \Theta) = 1 - \beta. \tag{7}$$

The distribution of $\mathscr{E}$ described in (7) encodes the expert opinion about θ, without additional assumptions about the distribution of values inside or outside B.

Now, besides the data-dependent random set $\Theta_y(\mathbb{S})$, the opinion encoder random set $\mathscr{E}$ also carries information about θ. We assume here that $\Theta_y(\mathbb{S})$ and $\mathscr{E}$ are independent, and the goal is combining these two sources of information to strengthen the conclusions that can be drawn about θ. However, it would not make sense if this additional information would ruin whatever validity properties the original inferential model has. Therefore, incorporation of the information in $\mathscr{E}$ to $\Theta_y(\mathbb{S})$ has to be done in a careful way, such that if data and expert opinions agree, then efficiency is gained and, regardless, the validity property of the inferential model is preserved.

A natural way combine two independent sources of information is through *Dempster's rule of combination* (e.g., Shafer, 1976; Kohlas and Monney, 1995), resulting in inferences that would be based on the conditional distribution

$$\Theta_y(\mathbb{S}) \cap \mathscr{E} \mid \Theta_y(\mathbb{S}) \cap \mathscr{E} \neq \varnothing$$

derived from the (product) joint distribution $\mathsf{P}_{\mathbb{S}} \times \mathsf{P}_{\mathscr{E}}$, where y is fixed at the observed data point. That is, Dempster's rule of combination considers a new random set which is the intersection of the two that carry information about θ, conditioning on the event $\{\Theta_y(\mathbb{S}) \cap \mathscr{E} \neq \varnothing\}$ of no conflict. It is through this conditioning step that "learning" takes place, sharpening inference. Ultimately, the plausibility function obtained from this combined random set is given by

$$p_y^{\mathrm{D}}(A) = 1 - \mathsf{P}_{\mathbb{S},\mathscr{E}}\{\Theta_y(\mathbb{S}) \cap \mathscr{E} \subseteq A^c \mid \Theta_y(\mathbb{S}) \cap \mathscr{E} \neq \varnothing\},$$

and the question is whether the inferential model corresponding to this combined plausibility function preserves the validity of the original. Specifically, is the distribution of $p_Y^{\mathrm{D}}(\{\theta\})$ stochastically no smaller than $\mathsf{Unif}(0,1)$ when $Y \sim \mathsf{P}_{Y|\theta}$?

Consider the normal example, $Y \sim \mathsf{P}_{Y|\theta} = \mathsf{N}(\theta, 1)$, where $B = [2,6]$, $\beta = 0.95$, and the true parameter takes values in $\{-1, 1, 4\}$. Figure 1 plots the distribution function, $\alpha \mapsto \mathsf{P}_{Y|\theta}\{p_Y^{\mathrm{D}}(\{\theta\}) \leq \alpha\}$, for each of the three values of θ, based on 10^4 Monte Carlo samples. Note that the distribution functions for $\theta = -1$ and $\theta = 1$, which correspond to cases where the prior is "misleading," are situated to the left of uniform, therefore, violating the validity property. The lack of validity is more extreme in the $\theta = 1$ case where the prior is "close to being right," which we expect to be the most common situation arising in practical applications. This makes sense, as $\Theta_y(\mathbb{S}) \cap B$ will likely have a significant intersection, and conditioning in this intersection being non-empty will tend to exclude θ more often than desired. On the other hand, if the prior is far from the truth, Dempster's rule of combination tends to down-weight the expert opinion, so that the plausibility function is effectively that based on $\Theta_y(\mathbb{S})$ alone, hence approximately valid. When the expert opinion is right, $p_Y^{\mathrm{D}}(\{\theta\})$ is both valid and efficient.

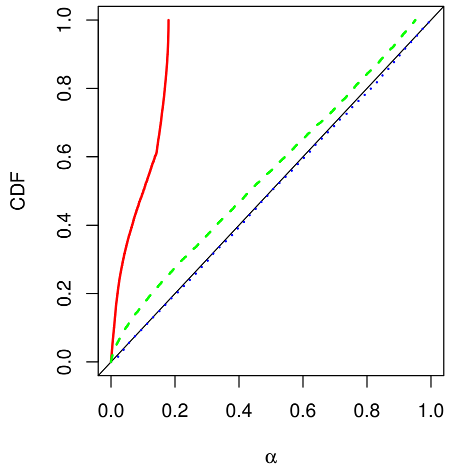


Figure 1: Distribution functions for plausibility contours based on Dempster's rule when $\theta = 1$ (red/solid), $\theta = -1$ (green/dashed), and $\theta = 4$ (blue/dotted), when $B = [2,6]$ and $\beta = 0.95$.

It should come as no surprise that Dempster's rule fails to preserve the validity property, given that the rule was developed without any sort of calibration properties in mind; but see (12). For us, preservation of the validity property is essential, so we seek an alternative combination rule under which validity can be maintained always and efficiency can be gained when expert opinion agrees with the data.

4. Valid Inferential Model With Expert Beliefs

4.1. Certain Case: $\beta = 1$

Recall that the random set $\mathbb{S}$ is designed to predict the unobserved value of the auxiliary variable U in $\mathbb{U}$ in such a way that validity and efficiency can be achieved. After $Y = y$ is observed, the constraint B can be mapped to a subset of $\mathbb{U}$:

$$\mathbb{U}_y(B) = \bigcup_{\vartheta \in B} \{u : y = a(\vartheta, u)\} \subseteq \mathbb{U}.$$

This subset includes all the possible u values that correspond to θ values in B with respect to the y that was observed. For the case where B is a constraint, the recommendation in Ermini Leaf and Liu (2012) was to consider an *elastic* random set $\mathbb{S}$ that would suitably stretch itself till it had non-empty intersection with $\mathbb{U}_y(B)$. This stretching-based strategy eliminates conflict cases but not via conditioning as Dempster's rule would do. Importantly, if $\mathbb{S}$ already leads to a valid inferential model, then stretching $\mathbb{S}$ will preserve the validity property, but causes some inefficiency due to the excessive overall size. To overcome this inefficiency, Ermini Leaf and Liu (2012) suggest to construct the inferential model using only the intersection of $\mathbb{U}_y(B)$ and the stretched $\mathbb{S}$, chopping off much of the excess size, and they show that this construction, too, satisfies the validity property.[1] But their strategy makes sense only when $\beta = 1$, i.e., when B is a constraint. Here we reformulate their developments in a way that will shed light on the more interesting case of $\beta \in (0,1)$ that is the focus of Section 4.2.

Our strategy will be to equip the random set $\mathbb{S}$ with two elasticity parameters, e and E, leading to a *dually elastic* random set, denoted by $\mathbb{S}_{e,E}$, where

- $e \geq 0$ is a *stretching* parameter that controls how far $\mathbb{S}$ is stretched toward $\mathbb{U}_y(B)$;
- and $E \geq 0$ is a *contraction* parameter that controls how much the stretched $\mathbb{S}$ gets shrunk in the opposite direction to improve efficiency.

For the normal example above, with $\mathbb{S}$ as in (5), a realization of $\mathbb{S}_{e,E}$ would take the form

$$[\min \mathbb{S} + E, \max \mathbb{S} + e] \quad \text{or} \quad [\min \mathbb{S} - e, \max \mathbb{S} - E], \tag{8}$$

depending on where the realization of $\mathbb{S}$ is situated relative to $\mathbb{U}_y(B)$. The question, of course, is how to select e and E.

Following Ermini Leaf and Liu (2012), we set e to be the smallest value at which the intersection of $\mathbb{S}_{e,E}$ and $\mathbb{U}_y(B)$ is non-empty, i.e.,

$$\hat{e} = \min\{e : \mathbb{S}_{e,E} \cap \mathbb{U}_y(B) \neq \varnothing\}.$$

Note that $\hat{e}$ does not depend on E. In the context of the normal example above, if the expert is certain that θ lies in the interval B, so that

$$\mathbb{U}_y(B) = [y - \max B, y - \min B],$$

then

$$\hat{e} = \begin{cases} \min \mathbb{U}_y(B) - \max \mathbb{S} & \text{if } \max \mathbb{S} < \min \mathbb{U}_y(B) \\ \min \mathbb{S} - \max \mathbb{U}_y(B) & \text{if } \min \mathbb{S} > \max \mathbb{U}_y(B) \\ 0 & \text{otherwise.} \end{cases}$$

For the contraction parameter, E, first recall that we are working with a random set which, in the absence of expert opinion, yields a valid and efficient inferential model. What characterizes the validity and efficiency of that inferential model is the plausibility contour of $\mathbb{S}$, given by $\gamma(u) = \mathsf{P}_{\mathbb{S}}(\mathbb{S} \ni u)$. If $\gamma(U) \sim \mathsf{Unif}(0,1)$, as a function of $U \sim P_U$, then the inferential model is both valid and efficient. Our proposal is to choose E in such a way that it preserves the original plausibility contour's properties even after the stretching by $\hat{e}$. That is, if

$$\gamma_E(u) = \mathsf{P}_{\mathbb{S}}(\mathbb{S}_{\hat{e},E} \ni u) \tag{9}$$

is the plausibility contour for the adjusted random set with stretching parameter $\hat{e}$ set according to the formulation above, and contraction according to E, then the goal is to specify $E = \hat{E}$ in such a way that $\gamma_{\hat{E}}(U) \sim \mathsf{Unif}(0,1)$ as a function of $U \sim \mathsf{P}_U$, for $\theta \in B$. It can be shown that this leads to a valid plausibility function, i.e.,

$$\begin{aligned} \mathsf{P}_{Y|\theta}\{p_Y(\theta) \leq \alpha\} &= \mathsf{P}_{Y|\theta}\{\mathsf{P}_{\mathbb{S}_{\hat{e},\hat{E}}}\{\Theta_Y(\mathbb{S}_{\hat{e},\hat{E}}) \ni \theta\} \leq \alpha\} \\ &= \mathsf{P}_{Y|\theta}\{\mathsf{P}_{\mathbb{S}_{\hat{e},\hat{E}}}\{\mathbb{S}_{\hat{e},\hat{E}} \ni a_\theta^{-1}(Y)\} \leq \alpha\} \\ &= \mathsf{P}_U\{\gamma_{\hat{E}}\left(a_\theta^{-1}(a(\theta,U))\right) \leq \alpha\} \\ &= \mathsf{P}_U\{\gamma_{\hat{E}}(U) \leq \alpha\} = \alpha. \end{aligned}$$

Following Ermini Leaf and Liu (2012), consider

$$\hat{E} = \min\{E : \mathbb{S}_{\hat{e},E} \subseteq \mathbb{U}_y(B)\}, \tag{10}$$

so that $\mathbb{S}_{\hat{e},\hat{E}}$ is just the intersection of the stretched $\mathbb{S}$ with $\mathbb{U}_y(B)$. Such choice is proven to lead to a valid plausibility function for every assertion $A \subset B$.

In the context of our normal example, we have

$$\gamma(u) = 2\{1 - F_n(|u|)\}.$$

Then the validity and efficiency of the resulting inferential model is, according to the results in Martin and Liu (2013), a consequence of the simple fact that $\gamma(U) \sim \mathsf{Unif}(0,1)$ when $U \sim \mathsf{P}_U = \mathsf{N}(0, n^{-1})$. We omit the details for the sake of space, but it is a relatively simple exercise to show that the contour function of $\mathbb{S}_{\hat{e},\hat{E}}$ ($\hat{E}$ as in 10) also has a uniform distribution like described above, for $\theta \in B$. The only challenge is that both $\hat{e}$ and $\hat{E}$ depend on data but, when Y is replaced by $\theta + U$, the θ cancels and the uniform distribution is easy to see. The validity proof in Ermini Leaf and Liu (2012) is consistent with our claim that the choice of $\hat{E}$ maintains the uniformity property.

Henceforth we will use the double-dagger symbol, ‡, to denote quantities based on the dually elastic modifications described above. For example, $\mathbb{S}^\ddagger = \mathbb{S}_{\hat{e},\hat{E}}$ and the corresponding plausibility function is

$$p_y^\ddagger(A) = 1 - \mathsf{P}_{\mathbb{S}}\{\Theta_y(\mathbb{S}^\ddagger) \subseteq A^c\} \quad A \subseteq B. \tag{11}$$

4.2. Less-Than-Certain Case: $\beta \in (0,1)$

Both the stretching and contraction parameters are chosen in an "extreme" way when the set B is a constraint. That

[1] When $\beta = 1$, the validity property is satisfied when the expert is correct, i.e., when θ is indeed in B. It is not hard to imagine a situation in which an expert exaggerates his/her confidence, but if they are given a less-than-certain option but they choose certain, then they have to live with whatever mistakes are made as a consequence.

is, $\hat{e}$ stretches $\mathbb{S}$ all the way to $\mathbb{U}_y(B)$ and $\hat{E}$ contracts the stretched random set down to the extent that it includes no points outside of $\mathbb{U}_y(B)$. But in the less-than-certain case where $\beta \in (0,1)$ and, hence, B is not a constraint, these extreme choices generally will not be satisfactory. Here we can follow the general strategy laid out in Section 4.1 but different $\hat{e}$ and $\hat{E}$ are needed in this more difficult setting.

In this more ambitious context, the challenge is balancing the expert's degree of certainty in B with how much (or how little) the data agrees with B. The key point is that, unlike in the constraint case above, here we do not assume that the expert is "right," it is indeed possible that θ lies outside of B. So, since we aim to achieve validity uniformly over the entire parameter space, the elasticity parameters need to be chosen more carefully. To guide our mathematical developments, we list the following desiderata.

1. *Data is sovereign.* More conflict between data and expert opinion makes the latter less influential.
2. *The size of β matters.* Larger β makes expert opinion more influential.
3. *Learning takes place.* Larger n makes expert opinion less influential.
4. *Validity everywhere.* The validity property is maintained, regardless of expert opinion, for all hypotheses in and out of B.
5. *No losses of efficiency, only gains.* If data and expert opinion agree, then there are efficiency gains, but no significant loss of efficiency if they do not agree.

Desiderata 1–3 are related to the incorporation of expert opinion into the original inferential model and, therefore, correspond to the stretching parameter. Desiderata 4–5 are related to the inferential model's properties and, therefore, pertain to the contraction parameter. For sure, these are ambitious goals, and our strategies to achieve them are considered next.

4.2.1. STRETCHING

The key to generalizing the extreme stretching in Section 4.1 is to incorporate the expert's belief probability β about the set B. More specifically, if $\beta = 1$ corresponds to extreme stretching, then we simply stretch less when $\beta < 1$. We would not want to stretch proportionally to β, however, because that would negatively affect Desiderata 1 and 3. Instead we want an "updated" version of β that takes into account how influential the expert opinion is. For this we suggest

$$\beta_y = \mathsf{P}_{\mathbb{S},\mathscr{E}}\{\mathscr{E} = B \mid \Theta_y(\mathbb{S}) \cap \mathscr{E} \neq \varnothing\}. \tag{12}$$

This is reminiscent of Dempster's rule of combination discussed above, as it is based on the conditional distribution of $\mathscr{E}$, given the event $\{\Theta_y(\mathbb{S}) \cap \mathscr{E} \neq \varnothing\}$, based on the (independent) joint distribution of $(\mathbb{S},\mathscr{E})$. This updated belief probability is influenced primarily on the extent to which data and expert opinion agree, as summarized by $\Theta_y(\mathbb{S})$ and $\mathscr{E}$, respectively. For example, if data and expert opinion agree, then β_y will tend to be large; but if data and expert opinion do not strongly agree, then β_y will be decreasing in n, due to the fact that $\Theta_y(\mathbb{S})$ becomes more precise as n increases. Therefore, if we set the stretching parameter, e, as

$$\hat{e} = \beta_y \cdot \min\{e : \mathbb{U}_y(B) \cap \mathbb{S}_{e,E} \neq \varnothing\}, \tag{13}$$

then we immediately achieve Desiderata 1–3.

In the normal example considered here, with $B = [a,b]$, it is easy to show that

$$\beta_y = \begin{cases} \frac{2\beta F_n(y-a)}{2\beta F_n(y-a)+(1-\beta)} & \text{if } y < a \\ \frac{2\beta F_n(b-y)}{2\beta F_n(b-y)+(1-\beta)} & \text{if } y > b \\ \beta & \text{otherwise,} \end{cases}$$

and this determines the stretching parameter $\hat{e}$.

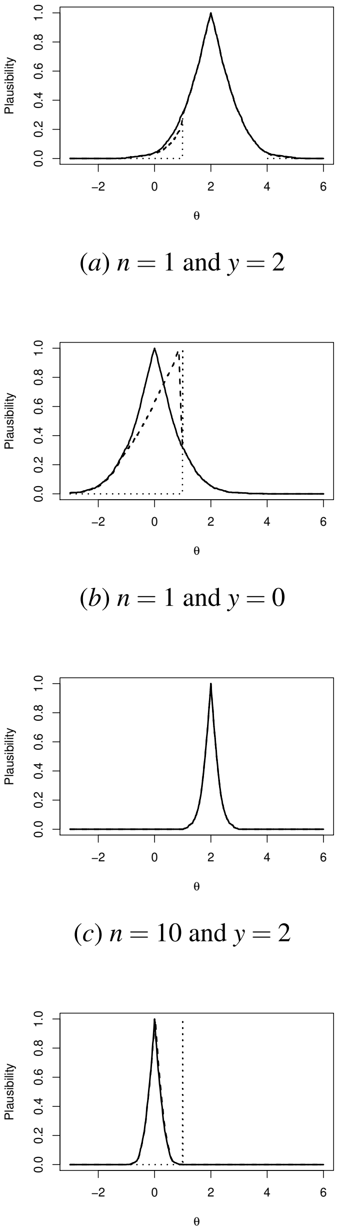
4.2.2. CONTRACTION

Stretching the original random set by any amount cannot damage the validity property. However, with the presence of expert opinion, there is an opportunity to gain in efficiency by a corresponding shrinking or contraction operation applied in the opposite direction of the stretching. But this contraction must be done carefully to ensure that efficiency is generally gained but validity is not lost, even if the expert opinion is "wrong."

To determine the contraction parameter, E, the idea of maintaining distributional properties of the original—without expert opinion—plausibility contour, described in Section 4.1, can be also applied here. Recall that $\gamma_E(u)$ in (9) is the plausibility contour for the adjusted random set with stretching parameter $\hat{e}$ and contraction parameter E. The goal is to specify $E = \hat{E}$ in such a way that $\gamma_{\hat{E}}(U) \sim \mathsf{Unif}(0,1)$ as a function of $U \sim \mathsf{P}_U$, but now not only for $\theta \in B$, but for θ since the entire parameter space is possible. The details of this derivation for our running normal example are in Appendix A.

5. Numerical Results

As an illustration, consider the normal mean example from Section 2. First, in Figure 2, we plot the plausibility contour $\vartheta \mapsto p_y^{\ddagger}(\{\vartheta\})$ for the inferential model based on the dually elastic random set $\mathbb{S}^{\ddagger}$ presented above; here we denote this as IM‡. The different settings we consider are $y \in \{0,2\}$ and $n \in \{1,10\}$. For the expert opinion, we take $B = [1,4]$ and consider two values of the belief probability, $\beta \in \{0.95, 1\}$. For comparison, we also display the plausibility contour for the original inferential model in Section 2 that ignores the expert's opinion. Panel (a) shows how we gain efficiency when the expert opinion agrees with the data; that is, the



(a) $n = 1$ and $y = 2$

(b) $n = 1$ and $y = 0$

(c) $n = 10$ and $y = 2$

(d) $n = 10$ and $y = 0$

Figure 2: Plots of the plausibility contour defined in (11), with $B = [1,4]$, $\beta = 1$ (dotted), and $\beta = 0.95$ (dashed). For comparison, we also display the plausibility contour for the basic inferential model in Section 2 that does not consider expert opinion (solid).

plausibility contour gets tighter when expert opinion is incorporated, clearly more so when $\beta = 1$ and the expert opinion is stronger. However, these gains mostly disappear in Panel (c) because the expert opinion is less influential when n is larger. In Panel (b) we can see how the plausibility contour accommodates the expert opinion when $\beta = 0.95$, but with virtually no loss of efficiency. Panel (d) shows that, with larger n, the data is more influential and pulls the plausibility contour away from the expert's opinion when $\beta = 0.95$; when $\beta = 1$, it is just a spike at $\vartheta = 1$.

Consider now a simulation study to compare the validity and efficiency of the proposed inferential model that incorporates expert opinion against that of several alternatives:

- the basic inferential model construction in Section 2 that ignores expert opinion;
- a Bayesian solution when a conjugate normal prior is chosen so that the prior probability is consistent with expert opinion, i.e., for given B and β, the prior hyperparameters are chosen so that it assigns roughly probability β to B;
- and an approach like described in Section 3 where expert opinion is encoded by $\mathscr{E}$ and combined with the basic inferential model via Dempster's rule.

Comparisons are made for $B = [2,9]$ and $B = [2,4]$, both with $\beta = 0.95$, and four values of the true parameter, $\theta \in \{3, 1.5, 0, -4\}$. Throughout, we take $n = 1$. For each scenario, 5000 data sets are generated and, from each, 95% plausibility/credible intervals are extracted and the coverage indicator and length is computed, resulting in coverage probability and mean length comparisons, displayed in Tables 1 and 2, respectively. We do not display the coverage probability for the basic inferential model's plausibility intervals because these are known to equal the target confidence level. Moreover, the mean length of these and of the Bayesian credible intervals are not shown because the lengths do not depend on data; for example, the 95% plausibility intervals for the basic inferential model always have length 3.92.

From Table 1, we can see that the target coverage probability is attained for all values of θ only for IM‡. The Bayesian solution is valid, or close to be valid, only when the prior is right, i.e., when θ is inside $[a,b]$. Dempster's approach is valid only when the prior is right or extremely wrong. Moreover, note how Dempster's approach performs quite poorly when the prior is close to being right, i.e., when $\theta = 1.5$.

From Table 2, we observe that gains in efficiency happen for IM‡ when B is right or close to being right. Moreover, more precision is gained when B is itself narrower. Finally, when $\theta = 0$, so that the expert opinion is somewhat misleading, we can observe a loss of efficiency in the plausibility intervals. But note that, when $\theta = -4$, so that the expert opinion is even more misleading, the lack of agreement between data and prior would make β_y in (12) close to zero and, hence, the IM‡ solution is effectively ignoring the

B	θ	Bayes	Dempster	IM‡
$[2,9]$	3.0	0.930	0.974	0.953
	1.5	0.828	0.594	0.948
	0.0	0.701	0.809	0.956
	−4.0	0.239	0.955	0.941
$[2,4]$	3.0	1.00	0.992	0.946
	1.5	0.080	0.601	0.957
	0.0	0.000	0.804	0.954
	−4.0	0.000	0.956	0.948

Table 1: Coverage probabilities 95% plausibility intervals for Bayes, Dempster's, and the IM‡ solutions with $\beta = 0.95$, $n = 1$, and various θ and B.

expert opinion and the standard solution is again obtained. Dempster's solution almost always provides narrower plausibility intervals, when compared to IM‡. However, such gains are only useful if the expert opinion is right or extremely wrong since, otherwise, the coverage probability is very low.

B	θ	Dempster	IM‡
$[2,9]$	3.0	3.02	3.89
	1.5	3.01	3.91
	0.0	4.01	3.95
	−4.0	3.90	3.92
$[2,4]$	3.0	2.18	3.85
	1.5	2.86	3.88
	0.0	4.02	3.95
	−4.0	3.89	3.92

Table 2: Mean length of 95% plausibility intervals for the Dempster and IM‡ approaches when $\beta = 0.95$, for different values of θ and B. All standard errors associated with the estimated averages are less than 0.05.

6. Conclusion

This paper focuses on the fundamentally important and challenging problem of combining data and vague prior information/expert opinion into an inferential model that provides valid and efficient inference about the parameter even if that prior input is wrong or misleading in some way. A Bayesian approach cannot accomplish this because (a) prior distributions require expert opinion at a higher resolution than is often available and (b) expert opinion that is misleading will introduce bias and negatively impact inferences. Dempster's rule, on the other hand, can easily accommodate the combination of data with vague prior information, but it does not preserve any validity properties that would hold when expert opinion is ignored. Here we deal with this problem by considering an inferential model driven by random sets with a dynamic, dually elastic behavior whereby they are stretched to accommodate expert opinion that more-or-less agrees with the data and, simultaneously, contracts the random set in the opposite direction just enough to preserve validity and achieve efficiency.

Our main contribution here is just the demonstration that a solution to this fundamental problem is possible. There are many technical details that remain to be sorted out. For one thing, we have only considered here the relatively simple normal mean problem. The same techniques can be applied under different statistical models and the performance is not affected; indeed, similarly promising results are obtained for Poisson data. But the more difficult question is about how to adjust the proposed stretching–contraction strategy when θ is multi-dimensional. Moreover, there are other kinds of expert opinions, e.g., about monotonicity (Altendorf et al., 2005; Feelders and van der Gaag, 2005), that cannot immediately be handled using the approach described here. Though there are still many unanswered questions, we are excited about this idea's potential and we will continue our investigation.

Acknowledgments

The authors are grateful for the valuable feedback provided by the three anonymous reviewers.

Appendix A. Contraction When $\beta \in (0,1)$

Let S_α be an arbitrary realization of the random set $\mathbb{S}$ that ignores the expert opinion such that

$$\mathsf{P}_U\{U \notin S_\alpha\} = \alpha.$$

To that same realization S_α, we can apply the stretching according to $\hat{e}$ in (13) and by some amount E, resulting in a new set $S_{\alpha,\hat{e},E}$ where

$$\mathsf{P}_U\{U \notin S_{\alpha,\hat{e},E}\} = \alpha_E.$$

The goal here is to choose $E = \hat{E}$ such that

1. $\alpha_{\hat{E}} = \alpha$ for all $\alpha \in (0,1)$;
2. the focal elements $S_{\alpha,\hat{e},\hat{E}}$ form a nested collection $\mathbb{S}_{\hat{e},\hat{E}}$.

Therefore, by such construction,

$$\mathsf{P}_{\mathbb{S}_{\hat{e},\hat{E}}}\{\mathbb{S}_{\hat{e},\hat{E}} \subseteq S_{\alpha,\hat{e},\hat{E}}\} \equiv \mathsf{P}_U\{U \in S_{\alpha,\hat{e},\hat{E}}\} = 1-\alpha.$$

Recall the plausibility contour, γ_E, in (9), and set $Q_E(u) = 1-\gamma_E(u)$. For $\hat{E}$ according to the construction above, note that $Q_{\hat{E}}(u) > 1-\alpha$ iff $u \in S^c_{\alpha,\hat{e},\hat{E}}$. Thus,

$$\mathsf{P}_U\{Q_{\hat{E}}(U) > 1-\alpha\} = \mathsf{P}_U\{U \in S^c_{\alpha,\hat{e},\hat{E}}\}$$

$$
\begin{aligned}
&= 1 - \mathsf{P}_U\{U \in S_{\alpha,\hat{e},\hat{E}}\} \\
&= \alpha.
\end{aligned}
$$

Since α is arbitrary, $Q_{\hat{E}}(U) \sim \mathsf{Unif}(0,1)$ and, consequently, $\gamma_{\hat{E}}(U) \sim \mathsf{Unif}(0,1)$.

The question now is how to obtain such $E = \hat{E}$. Consider the case where $\min S_\alpha < \min \mathbb{U}_y(B)$, so that we would want to contract to the right; the case of contracting to the left is analogous. The goal is to find $E = \hat{E}$ such that

$$\mathsf{P}_U\{U < \min S_\alpha + \hat{E}\} = \alpha - \lambda_\theta(\alpha),$$

where

$$\lambda_\theta(\alpha) = \mathsf{P}_U\{U > \max S_{\alpha,\hat{e},\hat{E}}\} \le \tfrac{\alpha}{2}.$$

Then $\hat{E}$, when treated as a constant, would be

$$\hat{E} = F_n^{-1}\big(\alpha - \lambda_\theta(\alpha)\big) - \min S_\alpha,$$

which depends on the unknown θ and, therefore, cannot be used. But this obstacle can be overcome by allowing $\hat{E}$ to depend on both α and the data y in an appropriate way. Then our specific proposal is to take $\hat{E} = y - \vartheta^\star - \min S_\alpha$, where $\vartheta^\star = \vartheta^\star(y,\alpha)$ solves

$$F_n^{-1}\big(\alpha - \lambda_\vartheta(\alpha)\big) = y - \vartheta.$$

This equation can be solved numerically using Monte Carlo to approximate the probabilities involved. And it is not difficult to show that the resulting focal elements form a nested collection, satisfying Condition 2 above, and also that

$$\mathsf{P}_U\{U < \min S_\alpha + \hat{E}\} = \alpha - \lambda_\theta(\alpha),$$

so condition 1 above also holds:

$$
\begin{aligned}
\alpha_{\hat{E}} &= \mathsf{P}_U\{U < \min S_\alpha + \hat{E}\} + \lambda_\theta(\alpha) \\
&= \alpha - \lambda_\theta(\alpha) + \lambda_\theta(\alpha) \\
&= \alpha.
\end{aligned}
$$

References

Eric Altendorf, Angelo C. Restificar, and Thomas G. Dietterich. Learning from sparse data by exploiting monotonicity constraints. In *Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence, UAI'05*, pages 18–26, 2005.

Michael S. Balch, Ryan Martin, and Scott Ferson. Satellite conjunction analysis and the false confidence theorem. `arXiv:1706.08565`, 2017.

Michael Scott Balch. Mathematical foundations for a theory of confidence structures. *Internat. J. Approx. Reason.*, 53(7):1003–1019, 2012.

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

Thierry Denœux and Shoumei Li. Frequency-calibrated belief functions: review and new insights. *Internat. J. Approx. Reason.*, 92:232–254, 2018.

Didier Dubois and Henri Prade. *Possibility Theory*. Plenum Press, New York, 1988.

Duncan Ermini Leaf and Chuanhai Liu. Inference about constrained parameters using the elastic belief method. *Internat. J. Approx. Reason.*, 53(5):709–727, 2012. ISSN 0888-613X.

Ad Feelders and Linda C. van der Gaag. Learning bayesian network parameters with prior knowledge about context-specific qualitative influences. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05, pages 193–200, 2005.

Ronald A. Fisher. *Statistical Methods and Scientific Inference*. Hafner Press, New York, 3rd edition, 1973.

Jan Hannig, Hari Iyer, Randy C. S. Lai, and Thomas C. M. Lee. Generalized fiducial inference: a review and new results. *J. Amer. Statist. Assoc.*, 111(515):1346–1361, 2016.

Jürg Kohlas and Paul-André Monney. *A Mathematical Theory of Hints*, volume 425 of *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, Berlin, 1995.

Ryan Martin. False confidence, non-additive beliefs, and valid statistical inference. `arXiv:1607.05051`, 2019.

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

Ryan Martin and Chuanhai Liu. *Inferential Models*, volume 147 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2016.

Ilya Molchanov. *Theory of Random Sets*. Probability and Its Applications (New York). Springer-Verlag London Ltd., London, 2005. ISBN 978-185223-892-3; 1-85233-892-X.

Hung T. Nguyen. *An Introduction to Random Sets*. Chapman & Hall/CRC, Boca Raton, FL, 2006. ISBN 978-1-58488-519-1; 1-58488-519-X.

Nancy Reid and David R. Cox. On some principles of statistical inference. *Int. Stat. Rev.*, 83(2):293–308, 2015.

Tore Schweder and Nils Lid Hjort. *Confidence, Likelihood, Probability*, volume 41 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, New York, 2016.

Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J., 1976.

Min-ge Xie and Kesar Singh. Confidence distribution, the frequentist distribution estimator of a parameter: a review. *Int. Stat. Rev.*, 81(1):3–39, 2013. ISSN 0306-7734.

S. L. Zabell. R. A. Fisher and the fiducial argument. *Statist. Sci.*, 7(3):369–387, 1992.