Validity-Preservation Properties of Rules for Combining Inferential Models

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

Nicholas Syring NASYRING@WUSTL.EDU
Department of Mathematics and Statistics, Washington University in St. Louis, USA

Abstract

An inferential model encodes the data analyst's degrees of belief about an unknown quantity of interest based on the observed data, posited statistical model, etc. Inferences drawn based on these degrees of belief should be reliable in a certain sense, so we require the inferential model to be *valid*. The construction of valid inferential models based on individual pieces of data is relatively straightforward, but how to combine these so that the validity property is preserved? In this paper we analyze some common combination rules with respect to this question, and we conclude that the best strategy currently available is one that combines via a certain dimension reduction step before the inferential model construction.

Keywords: belief function; conditioning; Dempster's rule; Dubois and Prade's rule; plausibility function; random set; statistical inference.

1. Introduction

Let Y_1 denote observable data, not necessarily scalar, and let there be a posited statistical model for Y_1 depending on a parameter $\theta \in \Theta$. Observed data, the statistical model, and perhaps other things, e.g., a Bayesian prior distribution, can be used to define a degrees of belief measure for the unknown parameter θ. This mapping from data, model, etc., to degrees of belief is called an *inferential model* and, as the name suggests, is intended to be used for statistical inference; see Section 2.1. Our notion of statistical inference is the process by which data informs an analyst's beliefs in assertions about the parameter θ. Although these beliefs are necessarily subjective—specific to the analyst—we agree with Reid and Cox (2015) that "calibrated inferences seem essential." Therefore, we insist that these beliefs satisfy an objective *validity* property (Section 2.2) which implies frequentist error rate control and, more generally, reliable inferences for scientific applications. Interestingly, validity cannot hold for additive degrees of belief, i.e., probabilities, so we focus exclusively here on non-additive beliefs. Construction of a valid inferential model, using random sets, was first proposed in Martin and Liu (2013) and then further developed in Martin and Liu (2016); a brief overview is given in Section 2.3.

Now suppose there is second piece of data, Y_2, independent of the first, whose distribution depends on the same unknown parameter θ. One can produce an inferential model for θ based on both Y_1 and Y_2 but, since they both carry information about θ, it makes sense to combine the two in some way. This paper investigates different combination strategies. A first combination strategy that comes to mind is Dempster's rule; see, e.g., Shafer (1976, Ch. 3) and Dempster (2008, 2014). Here, however, we demonstrate that Dempster's rule does not preserve the validity property. Next we consider a combination strategy coming from possibility theory (Dubois and Prade, 1988) and show that, while it does preserve validity in the examples we consider—and we conjecture that this validity-preservation holds more generally—it sacrifices in terms of efficiency. Therefore, since we insist on validity, and desire efficiency, some other combination strategy is needed.

The two combination strategies discussed above work by combining the Y_1- and Y_2-based inferential models. A different approach would be to carry out the combination step directly on the data, i.e., *before* constructing the individual inferential models, and we consider two such strategies in Section 4. There we show that an approach to combination based on dimension reduction is able to achieve both validity and efficiency. In Section 5 we apply this dimension reduction strategy in two examples, demonstrating the validity and efficiency of the resulting inferential model. Some concluding remarks are given in Section 6.

2. Inferential Models

2.1. Definition

For data Y and statistical model $\mathscr{M} = \{\mathsf{P}_{Y|\theta} : \theta \in \Theta\}$, consisting of a family of possible distributions for Y, indexed by a parameter θ, suppose the goal is to make inference on θ based on the observation $Y = y$. An *inferential model* is simply a map from these inputs (and possibly more) to a belief function defined on the parameter space Θ, i.e.,

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

with $b_y(\emptyset) = 0$ and $b_y(\Theta) = 1$. Then, $b_y(A)$ is interpreted as the data analyst's degrees of belief about the hypothesis

© 2019 R. Martin & N. Syring.

$A \subseteq \Theta$ based on data $Y = y$, relative to the posited model, etc. The above mapping is left vague so as to cover many familiar modes of inference within one inferential modeling framework: e.g., Bayesian inference, which includes a prior distribution for θ and updates according to Bayes's formula; generalized fiducial inference (e.g., Hannig et al., 2016), which includes a data-generating equation and norm; and the approach in Martin and Liu (2016) described below, which includes a data-generating equation and an appropriate random set. The distinguishing feature of the latter approach is that the direct incorporation of a random set ensures that the inferential model output is a non-additive belief function. Non-additivity turns out to be important in the context of statistical inference, as we discuss next.

2.2. Validity Property

It was recently shown in Balch et al. (2017) that inferential models that produce additive beliefs, i.e., probabilities, as output suffer from what is called *false confidence*. That is, there always exists false hypotheses to which the inferential model tends to assign large beliefs. So if one agrees with Reid and Cox that systematically misleading inferences should be avoided, as we do, then it is necessary to consider non-additive inferential models; see, also, Martin (2019). But non-additivity alone is not enough, some additional constraints are needed.

An inferential model is said to be *valid* if $\forall\, \alpha \in (0,1)$, $\forall\, A \subseteq \Theta$

$$\sup_{\theta \notin A} \mathsf{P}_{Y|\theta}\{b_Y(A) > 1-\alpha\} \leq \alpha. \tag{1}$$

In words, assigning high belief to false hypotheses—those that do not contain the true θ—is a rare event relative to the posited model. This prevents false confidence and, thereby, systematically misleading conclusions.

Define the dual function $p_y(A) = 1 - b_y(A^c)$. Here, following Shafer (1976), 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: $\forall\, \alpha \in (0,1)$, $\forall\, A \subseteq \Theta$

$$\sup_{\theta \in A} \mathsf{P}_{Y|\theta}\{p_Y(A) \leq \alpha\} \leq \alpha. \tag{2}$$

In words, this says that hypotheses which are not false—those that contain the true θ—will tend to be assigned relatively high plausibility.

The validity property offers a sort of calibration, so that it is clear what it means for a hypothesis to have "small" plausibility. It also immediately leads to decision procedures with guaranteed control on the frequentist error rates. In particular:

- the test that rejects $H_0 : \theta \in A$ if and only if $p_y(A) \leq \alpha$ has Type I error probability no more than α;
- and the set $\{\vartheta : p_y(\{\vartheta\}) > \alpha\}$ has coverage probability at least $1-\alpha$.

Our main goal is to produce valid inferential models, but since a valid inferential model is not necessarily unique, a notion of *efficiency* may help to identify a best valid model. For example, a hypothesis test which never rejects the null hypothesis is trivially valid, but clearly inefficient. Formally, we may think of efficiency as the extent to which the the outer inequality holds in (2). If this inequality actually holds with equality, then there is no loss of efficiency, whereas if this inequality is strict, then there is some loss of efficiency.

2.3. Constructing Valid Inferential Models

It turns out that some care is needed to construct a valid inferential model. To our knowledge, what follows is the only general construction.

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

$$Y = a(\theta, U), \quad U \sim \mathsf{P}_U, \tag{3}$$

where $U \in \mathbb{U}$ is an auxiliary variable and its distribution, P_U, is known*, i.e., 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*

$$\Theta_y(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \Theta_y(u). \tag{4}$$

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:

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

Under certain conditions on the user-specified $\mathcal{S}$, it can be shown that the inferential model constructed above achieves the validity property in (1). Indeed, the required condition is that the plausibility contour for $\mathcal{S}$ be calibrated to P_U in the sense that $\gamma(U) \geq_{\text{st}} \text{Unif}(0,1)$, as a function of $U \sim \mathsf{P}_U$, where $\gamma(u) = \mathsf{P}_{\mathcal{S}}(\mathcal{S} \ni u)$ is the coverage function. While this condition might be unfamiliar, it is actually relatively easy to arrange; in fact, the random sets in the examples that follow are all simple, and yet they satisfy the sufficient conditions of the aforementioned theorem. For more on this random set calibration property, see Chapters 4–5 of Martin and Liu (2016).

2.4. Simple Normal Example

Suppose data Y has distribution $\mathsf{P}_{Y|\theta} = \mathsf{N}(\theta, 1)$. Construction of a valid inferential model is straightforward in this case, but we give the details here for illustration.

A-step *Write $Y = \theta + U$, where $U \sim \mathsf{P}_U = \mathsf{N}(0,1)$, which is fully known. This yields $\Theta_y(u) = \{y - u\}$, a singleton.*

P-step *Since the target is an unobserved draw from $\mathsf{N}(0,1)$, it makes sense to introduce a random set that is symmetric around 0. Here we take*

$$\mathcal{S} = \{u : |u| \leq |\tilde{U}|\}, \quad \tilde{U} \sim \mathsf{N}(0,1).$$

Note that $\gamma(u) = 2\{1 - \Phi(|u|)\}$, so $\gamma(U)$ is exactly $\mathsf{Unif}(0,1)$ when $U \sim \mathsf{N}(0,1)$.

C-step *Putting the two previous pieces together gives*

$$\Theta_y(\mathcal{S}) = \{\vartheta : |y - \vartheta| \leq |\tilde{U}|\}, \quad \tilde{U} \sim \mathsf{N}(0,1). \tag{5}$$

Then, for example, the corresponding plausibility contours are given by

$$p_y(\{\vartheta\}) = 2\{1 - \Phi(|y - \vartheta|)\}, \tag{6}$$

where Φ is the standard normal distribution function.

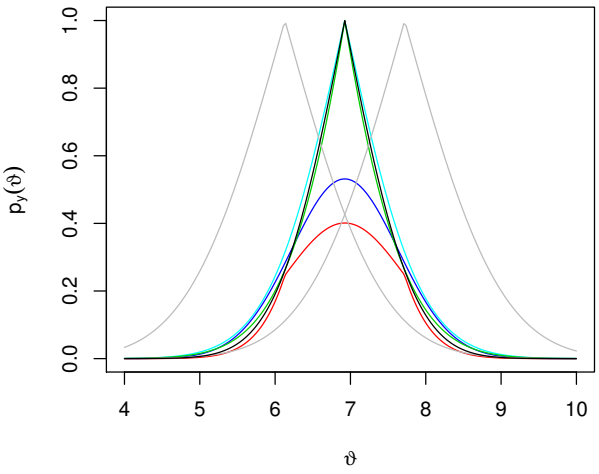

Figure 1: Plausibility contours based on the two individual observations, $y_1 = 6.13$ and $y_2 = 7.72$ (gray), along with those based on Dempster's combination rule (red); Dubois and Prade's rule (green, Section 3.2); the naive rules based on $r = 2$ (blue) and $r = \infty$ (cyan, Section 4.1); and the optimal rule (black, Section 4.2).

3. Combination Strategies

Suppose we have two independent data points, Y_1 and Y_2, both of which carry information about the same parameter θ; note that it is not necessary that the statistical models for the two be identical, e.g., in regression, the response variables have different distributions, due to the explanatory variables, but the parameters are the same in both models. We can construct valid inferential models for θ based on y_1 and y_2 individually using the strategy described above, but then how to combine?

3.1. Dempster's Rule

Dempster's rule says to combine by working with the intersection of the random sets defined by (4), conditioned on the event of no conflict. That is, Dempster's belief function $b_y^{\text{D}}(A)$ for the combined inferential model is

$$\mathsf{P}_{\mathcal{S}_1,\mathcal{S}_2}\Big\{\bigcap_{i=1}^2 \Theta_{y_i}(\mathcal{S}_i) \subseteq A \,\Big|\, \bigcap_{i=1}^2 \Theta_{y_i}(\mathcal{S}_i) \neq \varnothing\Big\}, \tag{7}$$

where $y = (y_1, y_2)$ and $\mathsf{P}_{\mathcal{S}_1,\mathcal{S}_2}$ is the joint distribution of $(\mathcal{S}_1, \mathcal{S}_2)$ under $\mathcal{S}_1 \sim \mathsf{P}_{\mathcal{S}_1}$ and $\mathcal{S}_2 \sim \mathsf{P}_{\mathcal{S}_2}$, assumed independent; see, e.g., Section 4.1 in Kohlas and Monney (1995).

The question here is if the two individual inferential models are valid, will combining them via Dempster's rule preserve their validity? Denœux and Li (2018) discuss the construction and application of belief functions with certain calibration properties but they do not investigate their combination via Dempster's rule. It turns out that Dempster's rule does not preserve validity, as the following example demonstrates.

Let Y_1 and Y_2 be independent and identically distributed (iid) $\mathsf{N}(\theta, 1)$, and write p_{y_i} for the plausibility function in (6) based on random set $\mathcal{S}_i$, $i = 1, 2$. To combine these two valid inferential models, we can insert the random sets (5) into the Dempster's rule formula (7). Here we write out the plausibility contour for the combined inferential model based on Dempster's formula. This is a rather tedious calculation but, in the end, for $y = (y_1, y_2)$ Dempster's plausibility contour $p_y^{\text{D}}(\{\vartheta\})$ equals

$$\frac{4\{1 - \Phi(|y_1 - \vartheta|)\}\{1 - \Phi(|y_2 - \vartheta|)\}}{1 - \{\Phi(2^{-1/2}|y_1 - y_2|) - \Phi(-2^{-1/2}|y_1 - y_2|)\}^2}.$$

To visualize this, Figure 1 plots this Dempster's rule-based combined plausibility contour, the two individual plausibility contours in (6), and a few others to be defined later. As expected, the peak of the combined plausibility contour is directly in between the two individual peaks, but the magnitude of the peak is much smaller, a consequence of the wide spread between the two data points in this case.

To the main question about validity of the combined inferential model, if validity did hold, then we would expect that $p_Y^{\text{D}}(\{\theta\})$, for the true θ, would have distribution stochastically no smaller than uniform. However, as we see

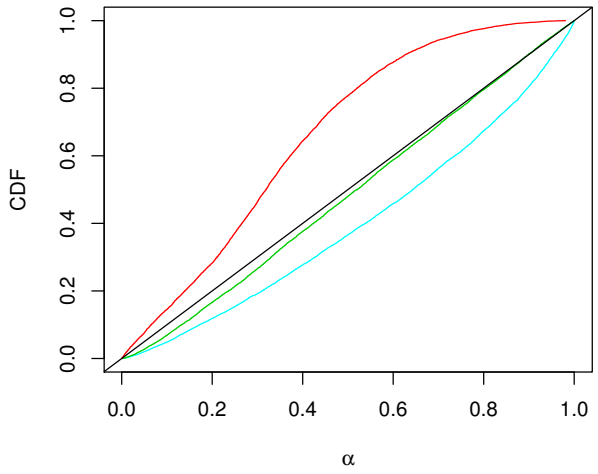

Figure 2: Distribution function $\alpha \mapsto \mathsf{P}_{Y|\theta}\{p_Y^{\bullet}(\{\theta\}) \leq \alpha\}$ for plausibility contours based on Dempster's rule (red), Dubois and Prade's rule (green), and the naive rule with $r = \infty$ (cyan), and the optimal (black); the distribution function for $r = 2$ is under the black line.

in Figure 2, based on 10^4 samples of $Y = (Y_1, Y_2)$, its distribution is situated considerably far to the left of uniform. Therefore, we can safely conclude that the combined inferential model, based on Dempster's rule, does not satisfy the validity property.

That Dempster's rule fails preserve the validity property is perhaps not surprising, since that rule was developed without any sort of calibration properties in mind. As the example above reveals, Dempster's rule yields plausibility contour values, $p_Y^{\text{D}}(\{\vartheta\})$ that tend to be too small. In particular, we desire $p_Y^{\text{D}}(\{\theta\})$ to be relatively large at the true θ, but Figure 2 shows that its distribution is concentrated in $[0.3, 0.6]$, with virtually no mass near 1. What would make the $p_Y^{\text{D}}(\{\theta\})$ values larger? Ideally, $\Theta_{y_1}(\mathcal{S}_1) \cap \Theta_{y_2}(\mathcal{S}_2)$ would contain θ for all/most realizations of $\mathcal{S}_1$ and $\mathcal{S}_2$, but this is not automatic. Indeed, since $\mathcal{S}_1$ and $\mathcal{S}_2$ are independent, it is possible to have one wide and one narrow, in which case $\Theta_{y_1}(\mathcal{S}_1) \cap \Theta_{y_1}(\mathcal{S}_2)$ misses $\bar{y}$ and, likely, θ too. Unfortunately, conditioning on the intersection being non-empty does not correct for this.

3.2. Dubois and Prade's Rule

Equation (63) in Dubois and Prade (1988) gives a formula for combining two (or more) possibility distributions. Since the two valid inferential models constructed above are based on nested random sets, the corresponding plausibility contours are possibility distributions, so we can combine them using Dubois and Prade's formula. For the case of two plausibility contours, the rule is relatively simple:

$$p_y^{\text{DP}}(\{\vartheta\}) = \frac{p_{y_1}(\{\vartheta\}) \wedge p_{y_2}(\{\vartheta\})}{\sup_t p_{y_1}(\{t\}) \wedge p_{y_2}(\{t\})},$$

where "$\wedge$" is the min operator. The idea behind this rule is that the minimum shrinks the individual plausibility contours towards a central region where both are relatively large; then the denominator is just a normalizer that makes the maximum value 1. In that normal example, this expression simplifies:

$$p_Y^{\text{DP}}(\{\vartheta\}) = \frac{p_{Y_1}(\{\vartheta\}) \wedge p_{Y_2}(\{\vartheta\})}{p_{Y_1}(\{\bar{Y}\})}, \tag{8}$$

where $\bar{Y} = (Y_1 + Y_2)/2$ is the average; in the denominator, Y_1 could be replaced by Y_2, both expressions would have the same numerical value. A plot of this function for a single pair of observations is shown in Figure 1 and its peak at $\bar{y}$ and tighter spread than the individual plausibility contours is clear.

Again the question is if this combination strategy can preserve the validity property of the individual inferential models being combined. It turns out that, at least for the normal example considered here, validity *is* preserved under Dubois and Prade's rule.

Let $(W_1, W_2) := (p_{Y_1}(\{\theta\}) \wedge p_{Y_2}(\{\theta\}), p_Y^{\text{DP}}(\{\theta\}))$, the numerator of the plausibility function and itself, at the true θ value. After some routine-but-tedious calculation, it can be shown that the joint density of (W_1, W_2) is given by

$$\frac{w_1}{w_2^2} \frac{\phi(\Phi^{-1}(1 - w_1/2) - 2\Phi^{-1}(1 - w_1/2w_2))}{\phi(\Phi^{-1}(1 - w_1/2w_2))},$$

for $0 \leq w_1 \leq w_2 \leq 1$ where $\phi = \Phi'$ denotes the standard normal density function. By numerically integrating first over w_1 and then over the region $W_2 \leq w_2$, we produce the distribution of $p_Y^{\text{DP}}(\{\theta\}))$ shown in Figure 2. Since the distribution function of $p_Y^{\text{DP}}(\{\theta\}))$ is strictly larger than a uniform, validity is preserved under combination, but the figure shows signs of inefficiency. It is possible for the plausibility values to be smaller, hence a narrower plausibility contour, without sacrificing validity. But this and Figure 1 both suggest that there may not be much room for improvement.

4. Alternative Combination Strategies

4.1. Naive Approach

It was mentioned above that nothing about the inferential model construction was specific to having only a single data point. So a naive strategy is to write the association as the pair/system of associations based on the individual data points. In that normal mean example in Section 3, this yields an auxiliary variable $U = (U_1, U_2)$, a pair of independent standard normals, to be predicted with a random set.

With only a single $\mathsf{N}(0,1)$ auxiliary variable, the symmetric interval random set was most natural, but with even just a pair, there are lots of reasonable choices. Two that we will consider here are

$$\mathcal{S} = \{u : \|u\|_r \leq \|\tilde{U}\|_r\}, \quad \tilde{U} \sim \mathsf{N}_2(0, I_2), \quad r = 2, \infty.$$

That is, we consider random sets shaped like circles and squares, both can be viewed as generalizations of the symmetric random interval used in Section 2.4. With a little bit of effort, the inferential models based on the circle and square random sets can be derived and their respective plausibility contours are given by

$$p_y^{\text{N}}(\{\vartheta\}) = \begin{cases} 1 - F_2(\|y - \vartheta 1_2\|_2^2) & r = 2 \\ \frac{1 - \{2\Phi(|y_1 - \vartheta| \vee |y_2 - \vartheta|) - 1\}^2}{1 - \{2\Phi(2^{-1}|y_1 - y_2|) - 1\}^2} & r = \infty, \end{cases}$$

where $y = (y_1, y_2)$ denotes the pair of observations, F_k denotes the chi-square distribution function with k degrees of freedom, and "$\vee$" is the max operator. It follows from the general theory in Martin and Liu (2016) that both are valid, but the two differ in very important ways.

To see how the two differ, first note that, for the $r = \infty$ case, Figure 2 shows that the distribution is to the right of $\mathsf{Unif}(0,1)$, which confirms validity, but suggests inefficiency. In the $r = 2$ case, $p_Y(\{\theta\})$ has exactly a $\mathsf{Unif}(0,1)$ distribution when θ is the true parameter, but its plausibility contour, displayed in Figure 1, still suggests inefficiency since its tails are wider than the optimal contour. To better understand this inefficiency, suppose that ϑ is inside the usual 95% confidence interval, i.e.,

$$\sqrt{2}|\bar{y} - \vartheta| < 1.96.$$

Then the plausibility contour for the $r = 2$ case satisfies

$$\begin{aligned} p_y^{\text{N}}(\{\vartheta\}) &= 1 - F_2(\|y - \bar{y}1_2\|_2^2 + 2|\bar{y} - \vartheta|^2) \\ &> 1 - F_2(\|y - \bar{y}1_2\|_2^2 + 1.96^2). \end{aligned}$$

In this case, $\|y - \bar{y}1_2\|_2^2$ tends to be close to 1, hence less than 1.96^2, which implies $p_y^{\text{N}}(\{\vartheta\}) > 0.05$. So we can conclude that the 95% plausibility interval, based on $r = 2$, contains the usual 95% confidence interval (with high probability), hence inefficiency. Figure 1 plots this plausibility contour ($r = 2$) against the "optimal" one, among others, and the former's wider spread is a consequence of the inefficiency just described.

4.2. A More Efficient Approach

The above inefficiency is a result of having to predict a two-dimensional auxiliary variable, U, for only a scalar parameter, θ. This suggests that efficiency can be gained by reducing the dimension, but how? The key insight, first discussed in Martin and Liu (2015), is that, when the auxiliary variable has higher dimension than the parameter, certain features of it are actually *observed*. Those observed features do not need to be predicted, hence an opportunity to effectively reduce the dimension. Moreover, by conditioning on the values of those observed features, prediction of the unobserved features can also be improved, hence even more efficiency gains.

In cases where there is information about a common parameter, θ, coming from multiple sources, it is possible to rewrite the baseline association, $Y = a(\theta, U)$, with $U \sim \mathsf{P}_U$ of dimension greater than that of θ, as

$$T_\theta(Y) = b(\theta, \tau(U)) \quad \text{and} \quad H(Y) = \eta(U),$$

for suitable one-to-one mappings $y \mapsto (T_\theta(y), H(y))$ and $u \mapsto (\tau(u), \eta(u))$; often T can be chosen free of θ, but see Section 5.2. The motivation for making such a change is that the second component does not involve θ, i.e., the feature $\eta(U)$ of U is *observed*, even though U is not. This implies that we only need to predict the lower-dimensional $\tau(U)$ with a random set and, moreover, we can condition on the observed value of $\eta(U)$ to sharpen those predictions. This process will be described below, after we discuss how to find the aforementioned mappings.

There are a number of ways to identify these mappings, the most common being based on sufficient statistics. However, in certain *non-regular* cases, like the example in Section 5.2, sufficiency may not lead to a satisfactory reduction in dimension, so some more sophisticated techniques are needed. Here we focus on an approach based on solving a suitable differential equation.

Start by noticing that the unobserved U be a solution $u_{y,\theta}$ to the equation $y = a(\theta, u)$. So in order for $\eta(U)$ to be observable, it must be that $\eta(u_{y,\theta})$ is not sensitive to changes in θ. In other words, it must be that

$$\frac{\partial \eta(u_{y,\theta})}{\partial \theta} = 0.$$

Of course, we can apply the chain rule to re-express this as

$$\frac{\partial \eta(u)}{\partial u}\Big|_{u = u_{y,\theta}} \cdot \frac{\partial u_{y,\theta}}{\partial \theta} = 0, \tag{9}$$

which is advantageous since the derivative of $u_{y,\theta}$ with respect to θ is often relatively straightforward. (The dimensions of the objects in the above display are left ambiguous here because they can vary from one context to another.) Sometimes it is possible to solve this equation via guess-and-check, in other cases more formal strategies, such as the method of characteristics (e.g., Polyanin et al., 2002), are needed. Once η is found, τ can be identified by the one-to-one constraint; similarly, H is determined by $H(y) = \eta(u_{y,\theta})$ and then T can be worked out too by its connection to H and τ.

Given the pair (τ, η), we make a change of auxiliary variable, to $V_1 = \tau(U)$ and $V_2 = \eta(U)$. Then the A-step

proceeds by specifying the set-valued map $\Theta_y(v_1) = \{\vartheta : T(y) = b(\vartheta, v_1)\}$. The relevant distribution is that of V_1, given $V_2 = h$, derived from P_U, where $h = H(y)$ is the value corresponding to the observed data $Y = y$. So, for the P-step, introduce a random set $\mathcal{S} = \mathcal{S}^{(h)}$ designed to predict a realization from $\mathsf{P}_{V_1|V_2=h}$. Finally, the C-step combines the A- and P-step results to get a new random set $\Theta_y(\mathcal{S}^{(h)})$ and gets a corresponding belief and plausibility function just like before,

$$b_y(A \mid h) = \mathsf{P}_{\mathcal{S}^{(h)}}\{\Theta_y(\mathcal{S}^{(h)}) \subseteq A, \} \quad \text{and}$$
$$p_y(A \mid h) = 1 - b_y(A^c \mid h).$$

As a quick illustration, consider the example with a pair of observations $Y = \theta 1_2 + U$. Solving for u gives $u_{y,\theta} = y - \theta 1_2$ and differentiating with respect to θ give $\partial u_{y,\theta}/\partial\theta = -1_2$, a constant vector. So we need η to be such that $\partial\eta(u)/\partial u$ is, say, a 2×2 matrix that sends constant vectors to 0. One option is to take $\eta(u) = Mu$, where $M = I_2 - 2^{-1}1_2 1_2^\top$ is the matrix that projects onto the space orthogonal to 1_2. Then the feature of U that is observed is the "residuals," i.e., $U - \bar{U}1_2 = Y - \bar{Y}1_2$. In the normal case, $V_1 = \tau(U) = \bar{U}$ is independent of the residuals, $\eta(U)$, so conditioning is unnecessary. The result is an inferential model based on $\bar{Y} = \theta + V_1$, the obvious choice based on sufficiency, which corresponds to the plausibility contour displayed in Figure 1.

5. Examples

5.1. Normal Fixed-Effects Model

Consider a generalization of the simple normal example above, namely, the fixed effects model where $Y = (Y_1, \ldots, Y_n)$ and $Y_i \sim \mathsf{N}(\theta, \sigma_i^2)$, independent, with σ_i known, $i = 1, \ldots, n$. Start with a baseline association

$$Y_i = \theta + \sigma_i U_i, \quad U_i \sim \mathsf{N}(0,1), \quad i = 1, \ldots, n.$$

From here, it is straightforward to follow the approach described in Section 2.4 and get plausibility contours, $p_{y_i}(\{\vartheta\})$, for each individual observation and then combine them according to Dubois and Prade's rule to get $p_y^{\text{DP}}(\{\vartheta\})$. See below for more on this.

Towards a potentially more efficient solution, follow the approach outlined in Section 4.2. That is, start by writing $u_{y,\theta} = \text{diag}(\omega)(y - \theta 1_n)$, where $\omega = (\sigma_1^{-1}, \ldots, \sigma_n^{-1})^\top$. Clearly, the derivative is $\partial u_{y,\theta}/\partial\theta = -\omega$, a constant vector, so a solution to the differential equation (9) is $\eta(u) = Mu$, where $M = I - \|\omega\|_2^{-2}\omega\omega^\top$ is a projection onto the space orthogonal to ω. A complementary function is $\tau(u) = \omega^\top u/\|\omega\|_2^2$, and the baseline association can be rewritten as

$$T(Y) = \theta + \tau(U) \quad \text{and} \quad H(Y) = \eta(U),$$

where $T(y) = \|\omega\|_2^{-2}\sum_{i=1}^n \omega_i^2 y_i$ and $H(y) = \text{diag}(\omega)(y - T(y)1_n)$. In this case, it is easy to see that $\tau(U)$ and $\eta(U)$ are independent—so conditioning has no effect—and $V = \tau(U)$ has a normal distribution with mean 0 and variance $\|\omega\|_2^{-2}$. Using the same random set as in Section 2.4, we arrive at a combined plausibility contour

$$p_y^\star(\{\vartheta\}) = 2\big\{1 - \Phi\big(\|\omega\|_2^{-1}|T(y) - \vartheta|\big)\big\}. \tag{10}$$

For a quick comparison of this combined inferential model compared to that based on Dubois and Prade's rule above, see Figure 3 for a plot of the plausibility contours for a single data set consisting of $n = 3$ observations with $\sigma = (1, 2, 4)$. Both combined plausibility contours hit roughly the center of the three observations, and have a narrower spread—both about the same—than the individual contours, a result of the information gained from combination. However, the peak of p_y^{DP} is at a different point than that of $p_y^\star$, but it turns out that the latter, $T(y)$, is the maximum likelihood estimator, which is optimal. Figure 4 plots the distribution function of $p_Y^{\text{DP}}(\{\theta\})$ and $p_Y^\star(\{\theta\})$, where θ is the true value, and again we see that Dubois and Prade's rule preserves validity but suffers some loss of efficiency.

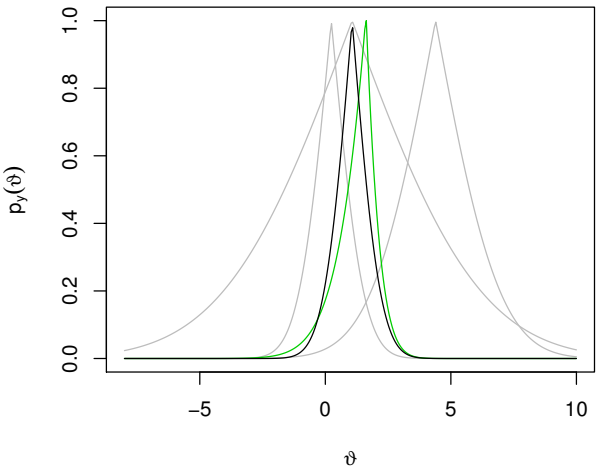

Figure 3: Plot of the plausibility contours for the $n = 3$ individual data points (gray), the combination based on Dubois and Prade's rule (green), and that in (10) for the normal fixed-effects model.

5.2. Curved Normal Model

Suppose we have n iid samples from a *curved normal* distribution, i.e., $\mathsf{N}(\theta, \theta^2)$, where the common θ in both the mean and variance is unknown; see, e.g., Searls (1964), Khan (1968), and Gleser and Healy (1976). This problem is *non-regular* in the sense that the minimal sufficient statistic is two-dimensional while the parameter is one-dimensional.

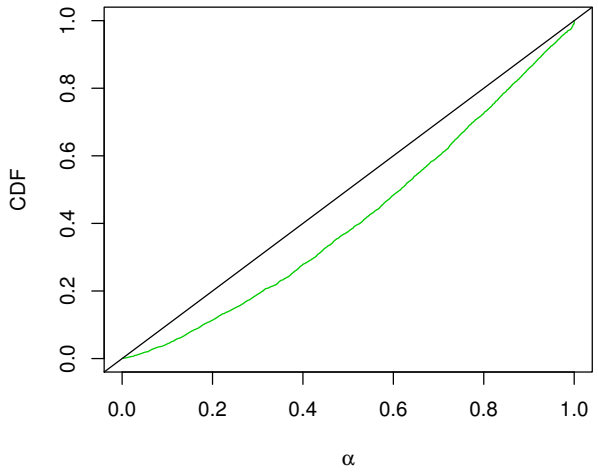

Figure 4: Plot of the distribution function $\alpha \mapsto \mathsf{P}_{Y|\theta}\{p_Y^{\bullet}(\{\theta\}) \leq \alpha\}$ for Dubois and Prade's rule (green) and the dimension reduction rule (black) in the normal fixed effects model.

Therefore, an estimator is necessarily not a sufficient statistic, so some information would be lost without some nontrivial adjustments. The usual strategy is to work with the conditional distribution of the estimator, given the value of an ancillary statistic (e.g., Fraser, 2004; Ghosh et al., 2010; Reid, 1995, 2003). The approach here, based on ideas in Section 4.2, is similar in spirit, but different in terms of both the rationale and the particular form of the solution. The more general version of this problem, as discussed in Brazauskas and Ghorai (2007), can be handled similarly.

Start with a baseline association in terms of the two-dimensional minimal sufficient statistic, $Y = (Y_1, Y_2)$, namely,

$$Y_1 = \theta + |\theta| U_1 \quad \text{and} \quad Y_2 = |\theta| U_2, \tag{11}$$

where Y_1 and Y_2 denote the sample mean and standard deviation, respectively, with $U_1 \sim \mathsf{N}(0, n^{-1})$ and $U_2^2 \sim (n-1)^{-1}\mathsf{ChiSq}(n-1)$, independent. For a given $y = (y_1, y_2)$, we have $u = u_{y,\theta} = |\theta|^{-1}(y_1 - \theta, y_2)^\top$, and differentiating with respect to θ gives

$$\frac{\partial u}{\partial \theta} = -\theta^{-1}\big(\text{sign}(\theta) + u_1, u_2\big)^\top,$$

where $\text{sign}(\theta) \in \{-1, +1\}$ denotes the sign of θ. Then it is easy to see that

$$\eta(u) = u_2^{-1}\big(\text{sign}(\theta) + u_1\big)$$

has a derivative that is orthogonal to $\partial u / \partial \theta$, hence is a candidate solution to our problem. It is slightly troubling that this η depends on the parameter, through $\text{sign}(\theta)$, but there are two ways to resolve this. Here, only for simplicity, we will assume that $\text{sign}(\theta)$ is *known*, which is not unreasonable, but one can alternatively employ the localization technique in Martin and Liu (2015, Sec. 6) that allows for certain dependence on the parameter.

To complete specification of the mappings, note that $\tau(u) = u_1/u_2$ complements η in the sense that, together, they make a one-to-one mapping. Moreover,

$$H(y) := \eta(u_{y,\theta}) = \frac{\text{sign}(\theta) + |\theta|^{-1}(y_1 - \theta)}{|\theta|^{-1} y_2} = \frac{y_1}{y_2},$$

the ratio of sample mean to variance. Finally, we take T_θ, depending on θ in this case, to be $T_\theta(y) = y_2^{-1}(y_1 - \theta)$.

Next, let $V_1 = \tau(U)$ and $V_2 = \eta(U)$; also write $h = y_1/y_2$ for the observed value of V_2. Then we need the conditional distribution of V_1, given $V_2 = h$, derived from the distribution of U described above. The calculation is somewhat tedious, but the density g_h for that conditional distribution is given by

$$\begin{aligned}\log g_h(v_1) = \text{const} &+ 3\log\big|\tfrac{1}{h - v_1}\big| - \tfrac{n}{2}\big(\tfrac{v_1}{h - v_1}\big)^2 \\ &+ (n-2)\log\big(\tfrac{\text{sign}(\theta)}{h - v_1}\big) - \tfrac{n-1}{2}\big(\tfrac{1}{h - v_1}\big)^2,\end{aligned}$$

where $\text{sign}(\theta)/(h - v_1) > 0$. We are now ready to complete the construction.

A-step *Write the dimension-reduced association as $T_\theta(Y) = V_1$, where $V_1 \sim g_h$. This defines the mapping $v_1 \mapsto \Theta_y(v_1) = \{\vartheta : T_\vartheta(y) = v_1\}$.*

P-step *For the random set that aims to predict the unobserved value of V_1, we suggest the highest-density region, namely,*

$$\mathcal{S}^{(h)} = \{v_1 : g_h(v_1) \geq g_h(\tilde{V}_1)\}, \quad \tilde{V}_1 \sim g_h,$$

C-step *Combine the two previous steps to get*

$$\Theta_y(\mathcal{S}^{(h)}) = \{\vartheta : g_h(T_\vartheta(y)) \geq g_h(\tilde{V}_1)\}, \quad \tilde{V}_1 \sim g_h,$$

and compute the belief and plausibility as discussed above. In particular, for $\vartheta > 0$ the plausibility contour is given by

$$p_y(\{\vartheta\}) = \mathsf{P}_{V_1|V_2=h}\{g_h(V_1) \leq g_h(T_\vartheta(y))\}. \tag{12}$$

A plot of the plausibility contours using the above approach and Dubois and Prade's formula based on a sample of size $n = 10$ from the curved normal distribution with $\theta = 2$ is shown in Figure 5. In this case, Dubois and Prade's rule combines the two separate plausibility contours based on the baseline association for the mean and standard deviation in (11). As a follow-up, we carried out a small simulation experiment to compare the performance of this inferential model to that based on a high-quality fiducial solution. We simulated 10,000 data sets of size $n = 10$

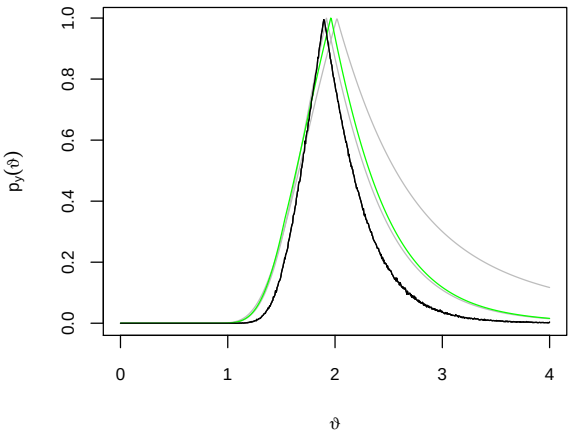

Figure 5: Plot of the plausibility contours in the curved normal example in Section 5.2, with $n = 10$ and $\theta = 2$ the true value: the plausibility contour in (12) (black), the contours based on the association in (11) (gray), and Dubois and Prade's rule combining the two (green).

with $\theta = 2$ as the true value and computed 95% confidence intervals for each method. Table 1 reports the estimated coverage probability and mean lengths of these intervals. As we can see, the inferential model hits the coverage probability target, a consequence of the validity property, while the fiducial solution falls a bit short. As expected, in order to achieve the target coverage probability, our proposed intervals have to be a bit longer, but this is a small price to pay for provable coverage claims.

Table 1: Coverage probability and mean length of interval estimates derived from the generalized fiducial approach in Pal Majumdar and Hannig (2015) compared with that of the valid inferential model constructed in Section 5.2.

	Coverage probability	Mean length
Fiducial	0.932	1.45
IM	0.946	1.56

6. Conclusion

The first part of this paper considered the combination of valid inferential models via the rules of Dempster and Dubois and Prade. The conclusion was that, unfortunately, Dempster's rule did not preserve validity while Dubois and Prade's rule did preserve validity but with a slight loss of efficiency. This, of course, is not an exhaustive list of the available combination rules, and future work will not only investigate the validity-preservation properties of Dubois and Prade's rule but also dig deeper into the belief function combination literature for alternative rules; we thank the reviewers for suggesting Daniel (2000), Smaradanche and Dezert (2006), Smets and Kennes (1994), and Yager (1987).

We went on to consider different combination strategies based on auxiliary variable dimension reduction, which allows one to achieve both validity and efficiency. At a high level, the main difference between the two classes of combination rules considered here is that the first carry out the combination *after* the individual inferential models have been constructed, whereas the second class does the combination *before* the inferential model construction. An interesting question, to be explored elsewhere, is if there exists a combination rule, operating on the individual inferential models themselves, not on the raw data, that can achieve both validity and efficiency. This is important because, in certain cases, such as meta-analysis applications, one may not have access to the raw data, only the inferential models reported from the individual analyses.

References

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

Vytaras Brazauskas and Jugal Ghorai. Estimating the common parameter of normal models with known coefficients of variation: a sensitivity study of asymptotically efficient estimators. *J. Stat. Comput. Simul.*, 77(7-8): 663–681, 2007. ISSN 0094-9655.

Milas Daniel. Distribution of contradictive belief masses in combination of belief functions. In B. Bouchon-Meunier, R. R. Yager, and L. A. Zadeh, editors, *Uncertainty and Fusion*, pages 431–446. Kluwer Academic Publishers, 2000.

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

A. P. Dempster. Statistical inference from a Dempster–Shafer perspective. In Xihong Lin, Christian Genest, David L. Banks, Geert Molenberghs, David W. Scott, and Jane-Ling Wang, editors, *Past, Present, and Future*

of Statistical Science, chapter 24. Chapman & Hall/CRC Press, 2014.

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

D. Dubois and H. Prade. Representation and combination of uncertainty with belief functions and possibility measures. *Comput. Intell.*, 4(3):244–264, 1988.

D. A. S. Fraser. Ancillaries and conditional inference. *Statist. Sci.*, 19(2):333–369, 2004. ISSN 0883-4237. With comments and a rejoinder by the author.

M. Ghosh, N. Reid, and D. A. S. Fraser. Ancillary statistics: A review. *Statist. Sinica*, 20:1309–1332, 2010.

Leon Jay Gleser and John D. Healy. Estimating the mean of a normal distribution with known coefficient of variation. *J. Amer. Statist. Assoc.*, 71(356):977–981, 1976. ISSN 0162-1459.

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.

Rasul A. Khan. A note on estimating the mean of a normal distribution with known coefficient of variation. *J. Amer. Statist. Assoc.*, 63(323):1039–1041, 1968.

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

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

A. Pal Majumdar and J. Hannig. Higher order asymptotics of generalized fiducial distributions. *Unpublished manuscript*, 2015.

A. D. Polyanin, V. F. Zaitsev, and A. Moussiaux. *Handbook of First Order Partial Differential Equations*, volume 1 of *Differential and Integral Equations and Their Applications*. Taylor & Francis Ltd., London, 2002. ISBN 0-415-27267-X.

N. Reid. The roles of conditioning in inference. *Statist. Sci.*, 10(2):138–157, 1995. ISSN 0883-4237.

N. Reid. Asymptotics and the theory of inference. *Ann. Statist.*, 31(6):1695–1731, 2003. ISSN 0090-5364.

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

Donald T. Searls. The utilization of a known coefficient of variation in the estimation procedure. *J. Amer. Statist. Assoc.*, 59(308):1225–1226, 1964.

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

F. Smaradanche and J. Dezert. Proportional conflict redistribution rules for information fusion. In *Advances and Applications of DSmT*, volume 2, chapter 1. American Research Press, 2006.

Philippe Smets and Robert Kennes. The transferable belief model. *Artificial Intelligence*, 66(2):191–234, 1994. ISSN 0004-3702.

Ronald R. Yager. On the Dempster-Shafer framework and new combination rules. *Inform. Sci.*, 41(2):93–137, 1987. ISSN 0020-0255.