

# Dempster–Shafer Theory and Statistical Inference with Weak Beliefs

Ryan Martin, Jianchun Zhang and Chuanhai Liu

*Abstract.* The Dempster–Shafer (DS) theory is a powerful tool for probabilistic reasoning based on a formal calculus for combining evidence. DS theory has been widely used in computer science and engineering applications, but has yet to reach the statistical mainstream, perhaps because the DS belief functions do not satisfy long-run frequency properties. Recently, two of the authors proposed an extension of DS, called the weak belief (WB) approach, that can incorporate desirable frequency properties into the DS framework by systematically enlarging the focal elements. The present paper reviews and extends this WB approach. We present a general description of WB in the context of inferential models, its interplay with the DS calculus, and the maximal belief solution. New applications of the WB method in two high-dimensional hypothesis testing problems are given. Simulations show that the WB procedures, suitably calibrated, perform well compared to popular classical methods. Most importantly, the WB approach combines the probabilistic reasoning of DS with the desirable frequency properties of classical statistics.

*Key words and phrases:* Bayesian, belief functions, fiducial argument, frequentist, hypothesis testing, inferential model, nonparametrics.

## 1. INTRODUCTION

A statistical analysis often begins with an iterative process of model-building, an attempt to understand the observed data. The end result is what we call a *sampling model*—a model that describes the data-generating mechanism—that depends on a set of unknown parameters. More formally, let  $X \in \mathbb{X}$  denote the observable data, and  $\Theta \in \mathbb{T}$  the parameter of interest. Suppose the sampling model  $X \sim P_{\Theta}$  can be rep-

resented by a pair consisting of (i) an equation

$$(1.1) \quad X = a(\Theta, U),$$

where  $U \in \mathbb{U}$  is called the *auxiliary variable*, and (ii) a probability measure  $\mu$  defined on measurable subsets of  $\mathbb{U}$ . We call (1.1) the *a-equation*, and  $\mu$  the *pivotal measure*. This representation is similar to that of Fraser [11], and familiar in the context of random data generation, where a random draw  $U \sim \mu$  is mapped, via (1.1), to a variable  $X$  with the prescribed distribution depending on known  $\Theta$ . For example, to generate a random variable  $X$  having an exponential distribution with fixed rate  $\Theta = \theta$ , one might draw  $U \sim \text{Unif}(0, 1)$  and set  $X = -\theta^{-1} \log U$ . For inference, uncertainty about  $\Theta$  is typically derived directly from the sampling model, without any additional considerations. But Fisher [10] highlighted the fundamental difference between sampling and inference, suggesting that the two problems should be, somehow, kept separate. Here we take a new approach in which inference is not determined by the sampling model alone—a so-called *inferential model* is built to handle posterior uncertainty separately.

---

Ryan Martin is Assistant Professor, Department of Mathematical Sciences, Indiana University–Purdue University Indianapolis, 402 North Blackford Street, Indianapolis, Indiana 46202, USA (e-mail: [rgmartin@math.iupui.edu](mailto:rgmartin@math.iupui.edu)). Jianchun Zhang is Ph.D. Candidate, Department of Statistics, Purdue University, 250 North University Street, West Lafayette, Indiana 47907, USA (e-mail: [zhang10@stat.purdue.edu](mailto:zhang10@stat.purdue.edu)). Chuanhai Liu is Professor, Department of Statistics, Purdue University, 250 North University Street, West Lafayette, Indiana 47907, USA (e-mail: [chuanhai@stat.purdue.edu](mailto:chuanhai@stat.purdue.edu)).

Since the early 1900s, statisticians have strived for inferential methods capable of producing posterior probability-based conclusions with limited or no prior assumptions. In Section 2 we describe two major steps in this direction. The first major step, coming in the 1930s, was Fisher’s fiducial argument, which uses a “pivotal quantity” to produce a posterior distribution with no prior assumptions on the parameter of interest. Limitations and inconsistencies of the fiducial argument have kept it from becoming widely accepted. A second major step, made by Dempster in the 1960s, extended both Bayesian and fiducial inference. Dempster uses (1.1) to construct a probability model on a class of subsets of  $\mathbb{X} \times \mathbb{T}$  such that conditioning on  $\Theta$  produces the sampling model, and conditioning on the observed data  $X$  generates a set of upper and lower posterior probabilities for the unknown parameter  $\Theta$ . Dempster [6] argues that this uncertainty surrounding the exact posterior probability is not an inconvenience but, rather, an essential component of the analysis. In the 1970s, Shafer [18] extended Dempster’s calculus of upper and lower probabilities into a general theory of evidence. Since then, the resulting Dempster–Shafer (DS) theory has been widely used in computer science and engineering applications but has yet to make a substantial impact in statistics. One possible explanation for this slow acceptance is the fact that the DS upper and lower probabilities are *personal* and do not satisfy the familiar long-run frequency properties under repeated sampling.

Zhang and Liu [25] have recently proposed a variation of DS inference that does have some of the desired frequency properties. The goal of the present paper is to review and extend the work of Zhang and Liu [25] on the theory of statistical inference with *weak beliefs* (WBs). The WB method starts with a belief function on  $\mathbb{X} \times \mathbb{T}$ , but before conditioning on the observed data  $X$ , a weakening step is taken whereby the focal elements are sufficiently enlarged so that some desirable frequency properties are realized. The belief function is weakened only enough to achieve the desired properties. This is accomplished by choosing a “most efficient” belief function from those which are sufficiently weak—this belief is called the *maximal belief* (MB) solution.

To emphasize the main objective of WB, namely, modifying belief functions to obtain desirable frequency properties, we present a new concept here called an *inferential model* (IM). Simply put, an IM is a belief function that is bounded from above by the conventional DS posterior belief function. For the special

case considered here, where the sampling model can be described by the a-equation (1.1) and the pivotal measure  $\mu$ , we consider IMs generated by using random sets to predict the unobserved value of the auxiliary variable  $U$ .

The remainder of the paper is organized as follows. Since WBs are built upon the DS framework, the necessary DS notation and concepts will be introduced in Section 2. Then, in Section 3, we describe the new approach to prior-free posterior inference based on the idea of IMs. Zhang and Liu’s WB method is used to construct an IM, completely within the belief function framework, and the desirable frequency properties of the resulting MB solution follow immediately from this construction. Sections 4 and 5 give detailed WB analyses of two high-dimensional hypothesis testing problems, and compare the MB procedures in simulations to popular frequentists methods. Some concluding remarks are made in Section 6.

## 2. FIDUCIAL AND DEMPSTER–SHAFFER INFERENCE

The goal of this section is to present the notation and concepts from DS theory that will be needed in the sequel. It is instructive, as well as of historical interest, however, to first discuss Fisher’s fiducial argument.

### 2.1 Fiducial Inference

Consider the model described by the a-equation (1.1), where  $\Theta$  is the parameter of interest,  $X$  is a sufficient statistic rather than the observed data, and  $U$  is the auxiliary variable, referred to as a pivotal quantity in the fiducial context. A crucial assumption underlying the fiducial argument is that each one of  $(X, \Theta, U)$  is uniquely determined by (1.1) given the other two. The pivotal quantity  $U$  is assumed to have an a priori distribution  $\mu$ , independent of  $\Theta$ . Prior to the experiment,  $X$  has a sampling distribution that depends on  $\Theta$ ; after the experiment, however,  $X$  is no longer a random variable. To produce a posterior distribution for  $\Theta$ , the variability in  $X$  prior to the experiment must somehow be transferred, after the experiment, to  $\Theta$ . As in Dempster [1], we “continue to believe” that  $U$  is distributed according to  $\mu$  after  $X$  is observed. This produces a distribution for  $\Theta$ , called the fiducial distribution.

EXAMPLE 1. To see the fiducial argument in action, consider the problem of estimating the unknown mean of a  $N(\Theta, 1)$  population based on a single observation  $X$ . In this case, we may write the a-equation (1.1) as

$$X = \Theta + \Phi^{-1}(U),$$

where  $\Phi(\cdot)$  is the cumulative distribution function (CDF) of the  $N(0, 1)$  distribution, and the pivotal quantity  $U$  has a priori distribution  $\mu = \text{Unif}(0, 1)$ . Then, for a fixed  $\theta$ , the fiducial probability of  $\{\Theta \leq \theta\}$  is, as Fisher [9] reasoned, determined by the following logical sequence:

$$\begin{aligned} \Theta \leq \theta &\iff X - \Phi^{-1}(U) \leq \theta \\ &\iff U \geq \Phi(X - \theta). \end{aligned}$$

That is, since the events  $\{\Theta \leq \theta\}$  and  $\{U \geq \Phi(X - \theta)\}$  are equivalent, their probabilities must be the same; thus, the fiducial probability of  $\{\Theta \leq \theta\}$ , as determined by “continuing to believe,” is  $\Phi(\theta - X)$ . We can, therefore, conclude that the fiducial distribution of  $\Theta$ , given  $X$ , is

$$(2.1) \quad \Theta \sim N(X, 1).$$

Note that (2.1) is exactly the objective Bayes answer when  $\Theta$  has the Jeffreys (flat) prior. A more general result along these lines is given by Lindley [15].

For a detailed account of the development of Fisher’s fiducial argument, criticisms of it, and a comprehensive list of references, see Zabell [24]. For more recent developments in fiducial inference, see Hannig [12].

## 2.2 Dempster–Shafer Inference

The Dempster–Shafer theory is both a successor of Fisher’s fiducial inference and a generalization of Bayesian inference. The foundations of DS have been laid out by Dempster [2–4, 6] and Shafer [18–22]. The DS theory has been influential in many scientific areas, such as computer science and engineering. In particular, DS has played a major role in the theoretical and practical development of artificial intelligence. The 2008 volume *Classic Works on the Dempster–Shafer Theory of Belief Functions* [23], edited by R. Yager and L. Liu, contains a selection of nearly 30 influential papers on DS theory and applications. For some recent statistical applications of DS theory, see Denoeux [7], Kohlas and Monney [13] and Edlefsen, Liu and Dempster [8].

DS inference, like Bayes, is designed to make probabilistic statements about  $\Theta$ , but it does so in a very different way. The DS posterior distribution is not a probability distribution on the parameter space  $\mathbb{T}$  in the usual (Bayesian) sense, but a distribution on a collection of subsets of  $\mathbb{T}$ . The important point is that a specification of an a priori distribution for  $\Theta$  is altogether avoided—the DS posterior comes from an a priori distribution over this collection of subsets of  $\mathbb{X} \times \mathbb{T}$  and

the DS calculus for combining evidence and conditioning on observed data.

Recall the a-equation (1.1) where  $X \in \mathbb{X}$  is the observed data,  $\Theta \in \mathbb{T}$  is the parameter of interest, and  $U \in \mathbb{U}$  is the auxiliary variable. In this setup,  $X$ ,  $\Theta$  and  $U$  are allowed to be vectors or even functions; the non-parametric problem where the parameter of interest is a CDF is discussed in Section 5. Here  $X$  is the full observed data and not necessarily a reduction to a sufficient statistic as in the fiducial context. Furthermore, unlike fiducial, the sets

$$(2.2) \quad \begin{aligned} \mathbb{T}_{x,u} &= \{\theta \in \mathbb{T} : x = a(\theta, u)\}, \\ \mathbb{U}_{x,\theta} &= \{u \in \mathbb{U} : x = a(\theta, u)\} \end{aligned}$$

are not required to be singletons.

Following Shafer [18], the key elements of the DS analysis are the *frame of discernment* and *belief function*; Dempster [6] calls these the *state space model* and the *DS model*, respectively. The frame of discernment is  $\mathbb{X} \times \mathbb{T}$ , the space of all possible pairs  $(X, \Theta)$  of real-world quantities. The belief function  $\text{Bel} : 2^{\mathbb{X} \times \mathbb{T}} \rightarrow [0, 1]$  is a set-function that assigns numerical values to events  $\mathcal{E} \subset \mathbb{X} \times \mathbb{T}$ , meant to represent the “degree of belief” in  $\mathcal{E}$ . Belief functions are generalizations of probability measures—see Shafer [18] for a full axiomatic development—and Shafer [20] shows that one can conveniently construct belief functions out of suitable measures and set-valued mappings through a “push-forward” operation. For our statistical inference problem, a particular construction comes to mind, which we now describe.

Consider the set-valued mapping  $M : \mathbb{U} \rightarrow 2^{\mathbb{X} \times \mathbb{T}}$  given by

$$(2.3) \quad M(U) = \{(X, \Theta) \in \mathbb{X} \times \mathbb{T} : X = a(\Theta, U)\}.$$

The set  $M(U)$  is called a *focal element*, and contains all those data-parameter pairs  $(X, \Theta)$  consistent with the model and particular choice of  $U$ . Let  $\mathcal{M} = \{M(U) : U \in \mathbb{U}\} \subseteq 2^{\mathbb{X} \times \mathbb{T}}$  denote the collection of all such focal elements. Then the mapping  $M(\cdot)$  in (2.3) and the pivotal measure  $\mu$  on  $\mathbb{U}$  together specify a belief function

$$(2.4) \quad \text{Bel}(\mathcal{E}) = \mu\{U : M(U) \subseteq \mathcal{E}\}, \quad \mathcal{E} \subset \mathbb{X} \times \mathbb{T}.$$

Some important properties of belief functions will be described below. Here we point out that  $\text{Bel}$  in (2.4) is the push-forward measure  $\mu M^{-1}$ , and this defines a probability distribution over measurable subsets of  $\mathcal{M}$ . Therefore, when  $U \sim \mu$ , one can think of  $M(U)$  as a

random set in  $\mathcal{M}$  whose distribution is defined by Bel in (2.4). Random sets will appear again in Section 3.

The rigorous DS calculus laid out in Shafer [18], and reformulated for statisticians in Dempster [6], makes the DS analysis very attractive. A key element of the DS theory is Dempster's rule of combination, which allows two (independent) pieces of evidence, represented as belief functions on the same frame of discernment, to be combined in a way that is similar to combining probabilities via a product measure. While the intuition behind Dempster's rule is quite simple, the general expression for the combined belief function is rather complicated and is, therefore, omitted; see Shafer [18], Chapter 3, or Yager and Liu [23], Chapter 1, for the details. But in a statistical context, the most important type of belief functions to be combined with Bel in (2.4) are those that fix the value of either the  $X$  or  $\Theta$  component—this type of combination is known as *conditioning*. It turns out that Dempster's rule of conditioning is fairly simple; see Theorem 3.6 of Shafer [18]. Next we outline the construction of these conditional belief functions, handling the two distinct cases separately.

*Condition on  $\Theta$ .* Here we combine the belief function (2.4) with another based on the information  $\Theta = \theta$ . Start with the trivial (constant) set-valued mapping

$$M_0(U) \equiv \{(X, \Theta) : \Theta = \theta\}.$$

This, together with the mapping  $M$  in (2.3), gives a combined focal element

$$M_0(U) \cap M(U) = \{(X, \theta) : X = a(\theta, U)\},$$

the  $\theta$ -cross section of  $M(U)$ , which we project down to the  $X$ -margin to give

$$(2.5) \quad M_\theta(U) = \{X : X = a(\theta, U)\} \subset \mathbb{X}.$$

Let  $\mathcal{A}$  be a measurable subset of  $\mathbb{X}$ . It can be shown that the conditional belief function  $\text{Bel}_\theta$  can be obtained by applying the same rule as in (2.4) but with  $M_\theta(U)$  in place of  $M(U)$ . That is, the conditional belief function, given  $\Theta = \theta$ , is given by

$$(2.6) \quad \begin{aligned} \text{Bel}_\theta(\mathcal{A}) &= \mu\{U : M_\theta(U) \subseteq \mathcal{A}\} \\ &= \mu\{U : a(\theta, U) \in \mathcal{A}\}, \end{aligned}$$

the push-forward measure defined by  $\mu$  and the mapping  $a(\theta, \cdot)$ , which is how the sampling distribution is defined. Therefore, given  $\Theta = \theta$ , the conditional belief function  $\text{Bel}_\theta(\cdot)$  is just the sampling distribution  $\text{P}_\theta(\cdot)$ .

*Condition on  $X$ .* For given  $X = x$ , we proceed just as before; that is, start with the trivial (constant) set-valued mapping

$$M_0(U) \equiv \{(X, \Theta) : X = x\}$$

and combine this with  $M(U)$  in (2.3) to obtain a new posterior focal element

$$M_0(U) \cap M(U) = \{(x, \Theta) : x = a(\Theta, U)\},$$

the  $x$ -cross section of  $M(U)$ , which we project down to the  $\Theta$  margin to give

$$(2.7) \quad M_x(U) = \{\Theta : x = a(\Theta, U)\} \subset \mathbb{T}.$$

Unlike the “condition on  $\Theta$ ” case above, this posterior focal element can, in general, be empty—a so-called *conflict case*. Dempster's rule of combination will effectively remove these conflict cases by conditioning on the event that  $M_x(U) \neq \emptyset$ ; see Dempster [3]. In this case, for an assertion, or hypothesis,  $\mathcal{A} \subset \mathbb{T}$ , the DS *posterior belief function*  $\text{Bel}_x$  is defined as

$$(2.8) \quad \text{Bel}_x(\mathcal{A}) = \frac{\mu\{U : M_x(U) \subseteq \mathcal{A}\}}{\mu\{U : M_x(U) \neq \emptyset\}}.$$

We now turn to some important properties of  $\text{Bel}_x$ . In Shafer's axiomatic development, belief functions are *nonadditive*, which implies

$$(2.9) \quad \text{Bel}_x(\mathcal{A}) + \text{Bel}_x(\mathcal{A}^c) \leq 1 \quad \text{for all } \mathcal{A},$$

with equality if and only if  $\text{Bel}_x$  is an ordinary additive probability. The intuition here is that evidence not in favor of  $\mathcal{A}^c$  need not be in favor of  $\mathcal{A}$ . If we define the *plausibility function* as

$$(2.10) \quad \text{Pl}_x(\mathcal{A}) = 1 - \text{Bel}_x(\mathcal{A}^c),$$

then it is immediately clear from (2.9) that

$$\text{Bel}_x(\mathcal{A}) \leq \text{Pl}_x(\mathcal{A}) \quad \text{for all } \mathcal{A}.$$

For this reason,  $\text{Bel}_x(\mathcal{A})$  and  $\text{Pl}_x(\mathcal{A})$  have often been called, respectively, the *lower* and *upper probabilities* of  $\mathcal{A}$  given  $X = x$ . In our statistical context,  $\mathcal{A}$  plays the role of a hypothesis about the unknown parameter  $\Theta$  of interest. So for any relevant assertion  $\mathcal{A}$ , the posterior belief and plausibility functions  $\text{Bel}_x(\mathcal{A})$  and  $\text{Pl}_x(\mathcal{A})$  can be calculated, and conclusions are reached based on the relative magnitudes of these quantities.

We have been writing “ $X = x$ ” to emphasize that the posterior focal elements and belief function are conditional on a fixed observed value  $x$  of  $X$ . But later we will consider sampling properties of the posterior belief function, for fixed  $\mathcal{A}$ , as a function of the random variable  $X$ , so, henceforth, we will write  $M_X(U)$  for  $M_x(U)$  in (2.7), and  $\text{Bel}_X$  for  $\text{Bel}_x$  in (2.8).

EXAMPLE 2. Consider again the problem in Example 1 of making inference on the unknown mean  $\Theta$  of a Gaussian population  $N(\Theta, 1)$  based on a single observation  $X$ . We can use the a-equation  $X = \Theta + \Phi^{-1}(U)$ , where  $U \sim \mu = \text{Unif}(0, 1)$ . The focal elements  $M(U)$  in (2.3) are the lines

$$M(U) = \{(X, \Theta) : X = \Theta + \Phi^{-1}(U)\}.$$

Given  $X$ , the focal elements  $M_X(U) = \{X - \Phi^{-1}(U)\}$  in (2.7) are singletons. Since  $U \sim \text{Unif}(0, 1)$ , the posterior belief function

$$\text{Bel}_X(\mathcal{A}) = \mu\{U : X - \Phi^{-1}(U) \in \mathcal{A}\}$$

is the probability that an  $N(X, 1)$  distributed random variable falls in  $\mathcal{A}$ , which is the same as the objective Bayes and fiducial posterior. Note also that this approach is different from that suggested by Dempster [2] and described in detail in Dempster [5].

EXAMPLE 3. Suppose that the binary data  $X = (X_1, \dots, X_n)$  consists of independent Bernoulli observations, and  $\Theta \in [0, 1]$  represents the unknown probability of success. Dempster [2] considered the sampling model determined by the a-equation

$$(2.11) \quad X_i = I_{\{U_i \leq \Theta\}}, \quad i = 1, \dots, n,$$

where  $I_A$  denotes the indicator of the event  $A$ , and the auxiliary variable  $U = (U_1, \dots, U_n)$  has pivotal measure  $\mu = \text{Unif}([0, 1]^n)$ . The belief function will have generic focal elements

$$M(U) = \{(X, \Theta) : X_i = I_{\{U_i \leq \Theta\}} \forall i = 1, \dots, n\}.$$

This definition of the focal element is quite formal, but looking more carefully at the a-equation (2.11) casts more light on the relationships between  $X_i$ ,  $U_i$  and  $\Theta$ . Indeed, we know that:

- if  $X_i = 1$ , then  $\Theta \geq U_i$ , and
- if  $X_j = 0$ , then  $\Theta < U_j$ .

Letting  $N_X = \sum_{i=1}^n X_i$  be the number of successes in the  $n$  Bernoulli trials, it is clear that exactly  $N_X$  of the  $U_i$ 's are smaller than  $\Theta$ , and the remaining  $n - N_X$  are greater than  $\Theta$ . There is nothing particularly important about the indices of the  $U_i$ 's, so throwing out conflict cases reduces the problem from the binary vector  $X$  and uniform variates  $U$  to the success count  $N = N_X$  and *ordered* uniform variates; see Dempster [2] for a detailed argument. Let  $U_{(i)}$  denote the  $i$ th order statistic from  $U_1, \dots, U_n$ , with  $U_{(0)} := 0$  and  $U_{(n+1)} := 1$ . Then the focal element  $M(U)$  above reduces to

$$M(U) = \{(N, \Theta) : U_{(N)} \leq \Theta \leq U_{(N+1)}\}, \\ U \in [0, 1]^n.$$

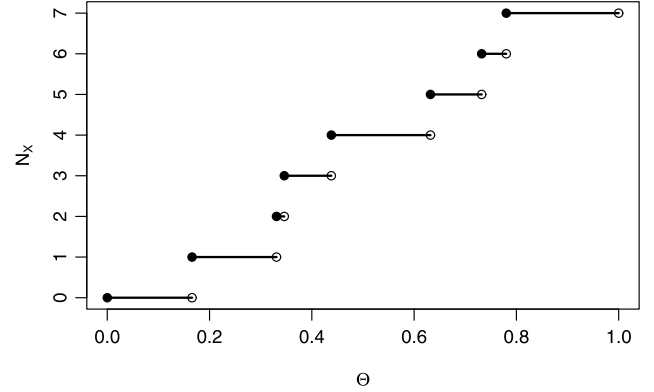


FIG. 1. A focal element  $M(U)$  for the Bernoulli data problem in Example 3, with  $n = 7$ . A posterior focal element is a horizontal line segment, the  $\Theta$ -interval determined by fixing the value of  $N = N_X$ .

Figure 1 gives a graphical representation of this generic focal element. Now given  $N$ , the posterior belief function has focal elements

$$(2.12) \quad M_N(U) = \{\Theta : U_{(N)} \leq \Theta \leq U_{(N+1)}\}, \\ U \in [0, 1]^n,$$

which are intervals (the horizontal lines in Figure 1) compared to the singletons in Example 2. Consider the assertion  $\mathcal{A}_\theta = \{\Theta \leq \theta\}$  for  $\theta \in [0, 1]$ . The posterior belief and plausibility functions for  $\mathcal{A}_\theta$  are given by

$$\text{Bel}_N(\mathcal{A}_\theta) = \mu\{U \in [0, 1]^n : U_{(N+1)} \leq \theta\}, \\ \text{Pl}_N(\mathcal{A}_\theta) = 1 - \mu\{U \in [0, 1]^n : U_{(N)} > \theta\}.$$

When  $N$  is fixed, the marginal beta distributions of  $U_{(N)}$  and  $U_{(N+1)}$  are available and  $\text{Bel}_X(\mathcal{A}_\theta)$  and  $\text{Pl}_X(\mathcal{A}_\theta)$  can be readily calculated. Plots for the case of  $n = 12$  and observed  $N = 7$  can be seen in Figure 3 (Example 5 in Section 3.3).

Next are two important remarks about the conventional DS analysis just described:

- The examples thus far have considered only “dull” assertions, such as  $\mathcal{A} = \{\Theta \leq \theta\}$ , where conventional DS performs fairly well. But for “sharp” assertions, such as  $\mathcal{A} = \{\Theta = \theta\}$ , particularly in high-dimensional problems, conventional DS can be too strong, resulting in plausibilities  $\text{Pl}_X(\mathcal{A}) \approx 0$  that are of no practical use.
- For fixed  $\mathcal{A}$ ,  $\text{Bel}_X(\mathcal{A})$  has no built-in long-run frequency properties as a function of  $X$ . Therefore, rules like “reject  $\mathcal{A}$  if  $\text{Pl}_X(\mathcal{A}) < 0.05$  or, equivalently, if  $\text{Bel}_X(\mathcal{A}^c) \geq 0.95$ ” have no guaranteed long-run error rates, so designing statistical *methodology* around conventional DS may be challenging.

It turns out that both of these problems can be taken care of by *shrinking*  $\text{Bel}_X$  in (2.8). We do this in Section 3 by suitably weakening the conventional DS belief, replacing the pivotal measure  $\mu$  with a belief function.

### 3. INFERENCE WITH WEAK BELIEFS

#### 3.1 Inferential Models

The conventional DS analysis of the previous section achieves the lofty goal of providing posterior probability-based inference without prior specification, but the difficulties mentioned at the end of Section 2.2 have kept DS from breaking into the statistical mainstream. Our basic premise is that these obstacles can be overcome by relaxing the crucial “continue to believe” assumption. The concept of *inferential models* (IMs) will formalize this idea.

Let  $\text{Bel}_X$  denote the posterior belief function (2.8) of the conventional DS analysis in Section 2.2, and let  $\text{Bel}^*$  be another belief function on the parameter space  $\mathbb{T}$ , possibly depending on  $X$ . For any assertion  $\mathcal{A}$  of interest,  $\text{Bel}^*(\mathcal{A})$  can be calculated and, at least in principle, used to make inference on the unknown  $\Theta$ . We say that  $\text{Bel}^*$  specifies an IM on  $\mathbb{T}$  if

$$(3.1) \quad \text{Bel}^*(\mathcal{A}) \leq \text{Bel}_X(\mathcal{A}) \quad \text{for all } \mathcal{A}.$$

Since  $\text{Bel}^*$  has plausibility  $\text{Pl}^*(\mathcal{A}) = 1 - \text{Bel}^*(\mathcal{A}^c)$ , it is clear from (3.1) that  $\text{Pl}^*(\mathcal{A}) \geq \text{Pl}_X(\mathcal{A})$  for all  $\mathcal{A}$ . Therefore, an IM can have meaningful nonzero plausibility even for sharp assertions. Shrinking the belief function can be done by suitably modifying the focal element mapping  $M(\cdot)$  or the pivotal measure  $\mu$ , but any other technique that generates a belief function bounded by  $\text{Bel}_X$  would also produce a valid IM.

$\text{Bel}_X$  itself specifies an IM, but is a very extreme case. At the opposite extreme is the vacuous belief function with  $\text{Bel}^*(\mathcal{A}) = 0$  for all  $\mathcal{A} \neq 2^{\mathbb{T}}$ . Clearly, neither of these IMs would be fully satisfactory in general. The goal is to choose an IM that falls somewhere in between these two extremes.

In the next subsection we use IMs to motivate the method of *weak beliefs*, due to Zhang and Liu [25]. That is, we apply their WB method to construct a particular class of IMs and, in Section 3.4, we show how a particular IM can be chosen.

#### 3.2 Weak Beliefs

Section 1 described how the a-equation might be used for data generation: fix  $\Theta$ , sample  $U$  from the pivotal measure  $\mu$ , and compute  $X = a(\Theta, U)$ . Now,

for the inference problem, suppose that the observed data  $X$  was, indeed, generated according to this recipe, but the corresponding values of  $\Theta$  and  $U$  remain hidden. Denote by  $U^*$  the value of the unobserved auxiliary variable; see (3.2). The key point is that knowing  $\Theta$  is equivalent to knowing  $U^*$ ; in other words, inference on  $\Theta$  is equivalent to *predicting* the value of the unobserved  $U^*$ . Both the fiducial and DS theories are based on this idea of shifting the problem of inference on  $\Theta$  to one of predicting  $U^*$ , although, to our knowledge, neither method has been described in this way before. The advantage of focusing on  $U^*$  is that the a priori distribution for  $U^*$  is fully specified by the sampling model.

More formally, if the sampling model  $\mathbf{P}_\Theta$  is specified by the a-equation (1.1), then the following relation must hold *after*  $X$  is observed:

$$(3.2) \quad X = a(\Theta, U^*),$$

where  $\Theta$  is unknown and  $U^*$  is unobserved. We can “solve” this equation for  $\Theta$  to get

$$(3.3) \quad \Theta \in A(X, U^*),$$

where  $A(\cdot, \cdot)$  is a set-valued map. Intuitively, (3.3) identifies those parameter values which are consistent with the observed  $X$ . For example, in the normal mean problem of Example 1, once  $X$  has been observed, there is a one-to-one relationship between the unknown mean  $\Theta$  and the unobserved  $U^*$ , that is,  $\Theta = A(X, U^*) = \{X - \Phi^{-1}(U^*)\}$ , so, given  $U^*$ , one can immediately find  $\Theta$ . Therefore, if we could predict  $U^*$ , then we could know  $\Theta$  exactly. The crucial “continue to believe” assumption of fiducial and DS says that  $U^*$  can be predicted by taking draws  $U$  from the pivotal measure  $\mu$ . WB weakens this assumption by replacing the draw  $U \sim \mu$  with a set  $\mathcal{S}(U)$  containing  $U$ , which is equivalent to replacing  $\mu$  with a belief function.

Recall from Section 2.2 that a measure and set-valued mapping together define a belief function. Here we fix  $\mu$  to be the pivotal measure, and construct a belief function on  $\mathbb{U}$  by choosing a set-valued mapping  $\mathcal{S}: \mathbb{U} \rightarrow 2^{\mathbb{U}}$  that satisfies  $U \in \mathcal{S}(U)$ . This is not the same as the DS analysis described in Section 2.2; there the belief function was fully specified by the sampling model, but here we must make a subjective choice of  $\mathcal{S}$ . We call this pair  $(\mu, \mathcal{S})$  a *belief*, as it generates a belief function  $\mu\mathcal{S}^{-1}$  on  $\mathbb{U}$ . Intuitively,  $(\mu, \mathcal{S})$  determines how aggressive we would like to be in predicting the unobserved  $U^*$ ; more aggressive means smaller  $\mathcal{S}(U)$ , and vice versa. We will call  $\mathcal{S}(U)$ , as a function of

$U \sim \mu$ , a *predictive random set* (PRS), and we can think of the inference problem as trying to hit  $U^*$  with the PRS  $\mathcal{S}(U)$ .

The two extreme IMs—the DS posterior belief function  $\text{Bel}_X$  in (2.8) and the vacuous belief function—are special cases of this general framework; take  $\mathcal{S}(U) = \{U\}$  for the former, and  $\mathcal{S}(U) = \mathbb{U}$  for the latter. So in this setting we see that the quality of the IM is determined by how well the PRS  $\mathcal{S}(U)$  can predict  $U^*$ . With this new interpretation, we can explain the comment at the end of Section 2.2 about the quality of conventional DS for sharp assertions in high-dimensional problems. Generally, high-dimensional  $\Theta$  goes hand-in-hand with high-dimensional  $U$ , and accurate estimates of  $\Theta$  require accurate prediction of  $U^*$ . But the *curse of dimensionality* states that, as the dimension increases, so too does the probabilistic distance between  $U^*$  and a random point  $U$  in  $\mathbb{U}$ . Consequently, the tiny (sharp) assertion  $\mathcal{A}$  will rarely, if ever, be hit by the focal elements  $M_X(U)$ .

In Section 3.4 we give a general WB framework, show how a particular  $\mathcal{S}$  can be chosen, and establish some desirable long-run frequency properties of the weakened posterior belief function. But first, in Section 3.3, we develop WB inference for given  $\mathcal{S}$  and give some illustrative examples.

### 3.3 Belief Functions and WB

In this section we show how to incorporate WB into the DS analysis described in Section 2.2. Suppose that a map  $\mathcal{S}$  is given. The case  $\mathcal{S}(U) = \{U\}$  was taken care of in Section 2.2, so what follows will be familiar. But this formal development of the WB approach will highlight two interesting and important properties, consequences of Dempster’s conditioning operation.

Previously, we have taken the frame of discernment to be  $\mathbb{X} \times \mathbb{T}$ . Here we have additional uncertainty about  $U^* \in \mathbb{U}$ , so first we will extend this to the larger frame  $\mathbb{X} \times \mathbb{T} \times \mathbb{U}$ . The belief function on  $\mathbb{U}$  has focal elements

$$\{U^* \in \mathbb{U} : U^* \in \mathcal{S}(U)\},$$

which correspond to cylinders in the larger frame, that is,

$$\{(X, \Theta, U^*) : U^* \in \mathcal{S}(U)\}.$$

Likewise, extend the focal elements  $M(U)$  in (2.3) to cylinders in the larger frame with focal elements

$$\{(X, \Theta, U^*) : X = a(\Theta, U^*)\}.$$

(The belief functions to which these extended focal elements correspond are implicitly formed by combining

the particular belief function with the vacuous belief function on the opposite margin.) Combining these extended focal elements, and simultaneously marginalizing over  $\mathbb{U}$ , gives a new focal element on the original frame  $\mathbb{X} \times \mathbb{T}$ , namely,

$$(3.4) \quad \begin{aligned} M(U; \mathcal{S}) &= \{(X, \Theta) : X = a(\Theta, u), u \in \mathcal{S}(U)\} \\ &= \bigcup \{M(u) : u \in \mathcal{S}(U)\}, \end{aligned}$$

where  $M(\cdot)$  is the focal mapping defined in (2.3). Immediately we see that the focal element  $M(U; \mathcal{S})$  in (3.4) is an expanded version of  $M(U)$  in (2.3). The measure  $\mu$  and the mapping  $M(U; \mathcal{S})$  generate a new belief function over  $\mathbb{X} \times \mathbb{T}$ :

$$\text{Bel}(\mathcal{E}; \mathcal{S}) = \mu\{U : M(U; \mathcal{S}) \subseteq \mathcal{E}\}.$$

Since  $M(U) \subseteq M(U; \mathcal{S})$  for all  $U$ , it is clear that  $\text{Bel}(\mathcal{E}; \mathcal{S}) \leq \text{Bel}(\mathcal{E})$ . The two DS conditioning operations will highlight the importance of this point.

*Condition on  $\Theta$ .* Conditioning on a fixed  $\Theta = \theta$ , the focal elements (as subsets of  $\mathbb{X}$ ) become

$$\begin{aligned} M_\theta(U; \mathcal{S}) &= \{X : X = a(\theta, u), u \in \mathcal{S}(U)\} \\ &= \bigcup \{M_\theta(u) : u \in \mathcal{S}(U)\}. \end{aligned}$$

This generates a new (predictive) belief function  $\text{Bel}_\theta(\cdot; \mathcal{S})$  that satisfies

$$\begin{aligned} \text{Bel}_\theta(\mathcal{A}; \mathcal{S}) &= \mu\{U : M_\theta(U; \mathcal{S}) \subseteq \mathcal{A}\} \\ &\leq \mu\{U : M_\theta(U) \subseteq \mathcal{A}\} \\ &= \text{Bel}_\theta(\mathcal{A}) = \text{P}_\theta(\mathcal{A}). \end{aligned}$$

Therefore, in the WB framework, this conditional belief function need not coincide with the sampling model as it does in the conventional DS context. But the sampling model  $\text{P}_\theta(\cdot)$  is *compatible* with the belief function  $\text{Bel}_\theta(\cdot; \mathcal{S})$  in the sense that

$$\text{Bel}_\theta(\cdot; \mathcal{S}) \leq \text{P}_\theta(\cdot) \leq \text{Pl}_\theta(\cdot; \mathcal{S}).$$

If we think about probability as a precise measure of uncertainty, then, intuitively, when we weaken our measure of uncertainty about  $U^*$  by replacing  $\mu$  with a belief function  $\mu\mathcal{S}^{-1}$ , we expect a similar smearing of our uncertainty about the value of  $X$  that will be ultimately observed.

*Condition on  $X$ .* Conditioning on the observed  $X$ , the focal elements (as subsets of  $\mathbb{T}$ ) become

$$\begin{aligned} M_X(U; \mathcal{S}) &= \{\Theta : X = a(\Theta, u), u \in \mathcal{S}(U)\} \\ &= \bigcup \{M_X(u) : u \in \mathcal{S}(U)\}. \end{aligned}$$

Evidently,  $M_X(U; \mathcal{S})$  is just an expanded version of  $M_X(U)$  in (2.7). But a larger focal element will be less likely to fall completely within  $\mathcal{A}$  or  $\mathcal{A}^c$ . Indeed, the larger  $M_X(U; \mathcal{S})$  generates a new posterior belief function  $\text{Bel}_X(\cdot; \mathcal{S})$  which satisfies

$$(3.5) \quad \begin{aligned} \text{Bel}_X(\mathcal{A}; \mathcal{S}) &= \mu\{U : M_X(U; \mathcal{S}) \subseteq \mathcal{A}\} \\ &\leq \mu\{U : M_X(U) \subseteq \mathcal{A}\} = \text{Bel}_X(\mathcal{A}). \end{aligned}$$

Therefore,  $\text{Bel}_X(\cdot; \mathcal{S})$  is a bonafide IM according to (3.1).

There are many possible maps  $\mathcal{S}$  that could be used. In the next two examples we utilize one relatively simple idea—using an interval/rectangle  $\mathcal{S}(U) = [A(U), B(U)]$  to predict  $U^*$ .

**EXAMPLE 4.** Consider again the normal mean problem in Example 1. The posterior belief function was derived in Example 2 and shown to be the same as the objective Bayes posterior. Here we consider a WB analysis where the set-valued mapping  $\mathcal{S} = \mathcal{S}_\omega$  is given by

$$(3.6) \quad \begin{aligned} \mathcal{S}(U) &= [U - \omega U, U + \omega(1 - U)], \\ \omega &\in [0, 1]. \end{aligned}$$

It is clear that the cases  $\omega = 0$  and  $\omega = 1$  correspond to the conventional and vacuous beliefs, respectively. Here we will work out the posterior belief function for  $\omega \in (0, 1)$  and compare the result to that in Example 2. Recall that the posterior focal elements in Example 2 were singletons  $M_X(U) = \{\Theta : \Theta = X - \Phi^{-1}(U)\}$ . It is easy to check that the weakened posterior focal elements are intervals of the form

$$\begin{aligned} M_X(U; \mathcal{S}) &= \bigcup \{M_X(u) : u \in \mathcal{S}(U)\} \\ &= [X - \Phi^{-1}(U + \omega(1 - U)), \\ &\quad X - \Phi^{-1}(U - \omega U)]. \end{aligned}$$

Consider the sequence of assertions  $\mathcal{A}_\theta = \{\Theta \leq \theta\}$ . We can derive analytical formulas for  $\text{Bel}_X(\mathcal{A}_\theta)$  and  $\text{Pl}_X(\mathcal{A}_\theta)$  as functions of  $\theta$ :

$$(3.7) \quad \begin{aligned} \text{Bel}_X(\mathcal{A}_\theta; \mathcal{S}) &= \left[1 - \frac{\Phi(X - \theta)}{1 - \omega}\right]^+, \\ \text{Pl}_X(\mathcal{A}_\theta; \mathcal{S}) &= 1 - \left[\frac{\Phi(X - \theta) - \omega}{1 - \omega}\right]^+, \end{aligned}$$

where  $x^+ = \max\{0, x\}$ . Plots of these functions are shown in Figure 2, for  $\omega \in \{0, 0.25, 0.5\}$ , when  $X = 1.2$  is observed. Here we see that as  $\omega$  increases, the spread between the belief and plausibility curves increases. Therefore, one can interpret the parameter  $\omega$  as a *degree of weakening*.

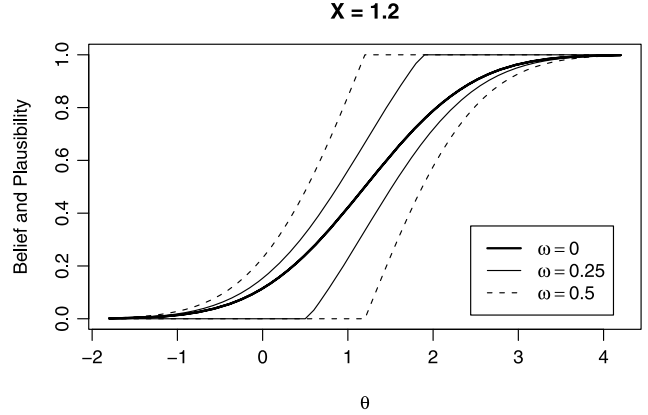


FIG. 2. Plots of belief and plausibility, as functions of  $\theta$ , for assertions  $\mathcal{A}_\theta = \{\Theta \leq \theta\}$  for  $X = 1.2$  and  $\omega \in \{0, 0.25, 0.5\}$  in the normal mean problem in Example 4. The case  $\omega = 0$  was considered in Example 1.

**EXAMPLE 5.** Consider again the Bernoulli problem from Example 3. In this setup, the auxiliary variable  $U = (U_1, \dots, U_n)$  in  $\mathbb{U} = [0, 1]^n$  is vector-valued. We apply a similar weakening principle as in Example 4, where we use a rectangle to predict  $U^*$ . That is, fix  $\omega \in [0, 1]$  and define  $\mathcal{S} = \mathcal{S}_\omega$  as

$$\mathcal{S}(U) = [A_1(U), B_1(U)] \times \dots \times [A_n(U), B_n(U)]$$

a Cartesian product of intervals like that in Example 4, where

$$\begin{aligned} A_i(U) &= U_i - \omega U_i, \\ B_i(U) &= U_i + \omega(1 - U_i). \end{aligned}$$

Following the DS argument in Example 3, it is not difficult to check that the (weakened) posterior focal elements are of the form

$$\begin{aligned} M_N(U; \mathcal{S}) &= [U_{(N)} - \omega U_{(N)}, U_{(N+1)} \\ &\quad + \omega(1 - U_{(N+1)})], \end{aligned}$$

an expanded version of the focal element  $M_X(U)$  in (2.12). Computation of the belief and plausibility can still be facilitated using the marginal beta distributions of  $U_{(N)}$  and  $U_{(N+1)}$ . For example, consider the sequence of assertions  $\mathcal{A}_\theta = \{\Theta \leq \theta\}$ ,  $\theta \in [0, 1]$ . Plots of  $\text{Bel}_N(\mathcal{A}_\theta; \mathcal{S})$  and  $\text{Pl}_N(\mathcal{A}_\theta; \mathcal{S})$ , as functions of  $\theta$ , are given in Figure 3 for  $\omega = 0$  (which is the conventional belief situation in Example 3) and  $\omega = 0.1$ , when  $n = 12$  and  $N = 7$ . As expected, the distance between the belief and plausibility curves is greater for the latter case. But this naive construction of  $\mathcal{S}$  is not the only approach; see Zhang and Liu [25] for a more efficient alternative based on a well-known relationship between the binomial and beta CDFs.



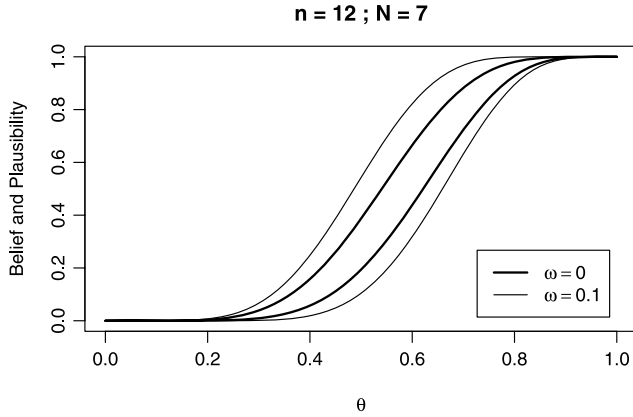


FIG. 3. Plots of belief and plausibility, as functions of  $\theta$ , for assertions  $\mathcal{A}_\theta = \{\Theta \leq \theta\}$  when  $n = 12$  and  $N = 7$  and  $\omega \in [0, 0.1]$ , in the Bernoulli success probability problem in Example 5. The case  $\omega = 0$  was considered in Example 3.

### 3.4 The Method of Maximal Belief

The WB analysis for a given set-valued map  $\mathcal{S}$  was described in Section 3.3. But how should one choose  $\mathcal{S}$  so that the posterior belief function satisfies certain desirable properties? Roughly speaking, the idea is to choose a map  $\mathcal{S}$  with the “smallest” PRSs  $\mathcal{S}(U)$  with the desired coverage probability. Following Zhang and Liu [25], we call this the method of *maximal belief* (MB).

Consider a general class of beliefs  $\mathcal{B} = (\mu, \mathcal{S})$ , where  $\mu$  is the pivotal measure from Section 1, and  $\mathcal{S} = \{\mathcal{S}_\omega : \omega \in \Omega\}$  is a class of set-valued mappings indexed by  $\Omega$ . Each  $\mathcal{S}_\omega$  in  $\mathcal{S}$  maps points  $u \in \mathbb{U}$  to subsets  $\mathcal{S}_\omega(u) \subset \mathbb{U}$  and, together with the pivotal measure  $\mu$ , determines a belief function  $\mu \mathcal{S}_\omega^{-1}$  on  $\mathbb{U}$  and, in turn, a posterior belief function  $\text{Bel}_X(\cdot; \mathcal{S}_\omega)$  on  $\mathbb{T}$  as in Section 3.3. For a given class of beliefs, it remains to choose a particular map  $\mathcal{S}_\omega$  or, equivalently, an index  $\omega \in \Omega$ , with the appropriate credibility and efficiency properties. To this end, define

$$(3.8) \quad Q_\omega(u) = \mu\{U : \mathcal{S}_\omega(U) \not\ni u\}, \quad u \in \mathbb{U},$$

which is the probability that the PRS  $\mathcal{S}_\omega(U)$  misses the target  $u \in \mathbb{U}$ . We want to choose  $\mathcal{S}_\omega$  in such a way that the random variable  $Q_\omega(U^*)$ , a function of  $U^* \sim \mu$ , is stochastically small.

DEFINITION 1. A belief  $(\mu, \mathcal{S}_\omega)$  is *credible* at level  $\alpha \in (0, 1)$  if

$$(3.9) \quad \varphi_\alpha(\omega) := \mu\{U^* : Q_\omega(U^*) \geq 1 - \alpha\} \leq \alpha.$$

Note the similarity between credibility and the control of Type-I error in the frequentist context of hypothesis testing. That is, if  $\mathcal{S}_\omega$  is credible at level  $\alpha = 0.05$ ,

then in a sequence of 100 similar inference problems, each having different  $U^*$ , we expect  $Q_\omega$ —the probability that the PRS  $\mathcal{S}_\omega$  misses its target—to exceed 0.95 in no more than 5 of these cases. The analogy with frequentist hypothesis testing is made here only to offer a way of understanding credibility.

It is not immediately clear why this notion of credibility is meaningful for the problem of inference on the unknown parameter  $\Theta$ . The following theorem, an extension of Theorem 3.1 in Zhang and Liu [25], gives conditions under which  $\text{Bel}_X(\cdot; \mathcal{S})$  has desirable long-run frequency properties in repeated  $X$ -sampling.

THEOREM 1. Suppose  $(\mu, \mathcal{S})$  is credible at level  $\alpha \in (0, 1)$ , and that  $\mu\{U : M_X(U; \mathcal{S}) \neq \emptyset\} = 1$ . Then, for any assertion  $\mathcal{A} \subset \mathbb{T}$ , the posterior belief function  $\text{Bel}_X(\mathcal{A}; \mathcal{S})$  in (3.5), as a function of  $X$ , satisfies

$$(3.10) \quad P_\Theta\{\text{Bel}_X(\mathcal{A}; \mathcal{S}) \geq 1 - \alpha\} \leq \alpha, \quad \Theta \in \mathcal{A}^c.$$

We can again make a connection to frequentist hypothesis testing, but this time in terms of assertions/hypotheses  $\mathcal{A}$  in the parameter space. If we adopt the decision rule “conclude  $\Theta \notin \mathcal{A}$  if  $\text{Pl}_X(\mathcal{A}; \mathcal{S}) < 0.05$ ,” then under the conditions of Theorem 1 we have

$$P_\Theta\{\text{Pl}_X(\mathcal{A}; \mathcal{S}) < 0.05\} \leq 0.05, \quad \Theta \in \mathcal{A}.$$

That is, if  $\mathcal{A}$  does contain the true  $\Theta$ , then we will “reject”  $\mathcal{A}$  no more than 5% of the time in repeated experiments, which is analogous to Type-I error probabilities in the frequentist testing domain. So the importance of Theorem 1 is that it equates credibility of the belief  $(\mu, \mathcal{S})$  to long-run error rates of belief/plausibility function-based decision rules. For example, the belief  $(\mu, \mathcal{S}_\omega)$  in (3.6) is credible for  $\omega \in [0.5, 1]$ , so decision rules based on (3.7) will have controlled error rates in the sense of (3.10). But remember that belief functions are posterior quantities that contain problem-specific evidence about the parameter of interest.

Credibility cannot be the only criterion, however since the belief, with  $\mathcal{S}(U) = \mathbb{U}$ , is always credible at any level  $\alpha \in (0, 1)$ . As an analogy, a frequentist test with empty rejection region is certain to control the Type-I error, but is practically useless; the idea is to choose from those tests that control Type-I error one with the largest rejection region. In the present context, we want to choose from those  $\alpha$ -credible maps the one that generates the “smallest” PRSs. A convenient way to quantify size of a PRS  $\mathcal{S}_\omega(U)$ , without using the geometry of  $\mathbb{U}$ , is to consider its coverage probability  $1 - Q_\omega$ .

DEFINITION 2.  $(\mu, \mathcal{S}_\omega)$  is as *efficient* as  $(\mu, \mathcal{S}_{\omega'})$  if

$$\varphi_\alpha(\omega) \geq \varphi_\alpha(\omega') \quad \text{for all } \alpha \in (0, 1).$$

That is, the coverage probability  $1 - Q_\omega$  is (stochastically) no larger than the coverage probability  $1 - Q_{\omega'}$ .

Efficiency defines a partial ordering on those beliefs that are credible at level  $\alpha$ . Then the level- $\alpha$  maximal belief ( $\alpha$ -MB) is, in some sense, the maximal  $(\mu, \mathcal{S}_\omega)$  with respect to this partial ordering. The basic idea is to choose, from among those credible beliefs, one which is most efficient. Toward this, let  $\Omega_\alpha \subset \Omega$  index those maps  $\mathcal{S}_\omega$  which are credible at level  $\alpha$ .

DEFINITION 3. For  $\alpha \in (0, 1)$ ,  $\mathcal{S}_{\omega^*}$  defines an  $\alpha$ -MB if

$$(3.11) \quad \varphi_\alpha(\omega^*) = \sup_{\omega \in \Omega_\alpha} \varphi_\alpha(\omega).$$

Such an  $\omega^*$  will be denoted by  $\omega(\alpha)$ .

By the definition of  $\Omega_\alpha$ , it is clear that the supremum on the right-hand side of (3.11) is bounded by  $\alpha$ . Under fairly mild conditions on  $\mathcal{S}$ , we show in Appendix A.1 that there exists an  $\omega^* \in \Omega_\alpha$  such that

$$(3.12) \quad \varphi_\alpha(\omega^*) = \alpha,$$

so, consequently,  $\omega^* = \omega(\alpha)$  specifies an  $\alpha$ -MB. We will, henceforth, take (3.12) as our working definition of MB. Uniqueness of a MB must be addressed case-by-case, but the left-hand side of (3.12) often has a certain monotonicity which can be used to show the solution is unique.

We now turn to the important point of computing the MB or, equivalently, the solution  $\omega(\alpha)$  of the equation (3.12). For this purpose, we recommend the use of a *stochastic approximation* (SA) algorithm, due to Robbins and Monro [17]. Kushner and Yin [14] give a detailed theoretical account of SA, and Martin and Ghosh [16] give an overview and some recent statistical applications.

Putting all the components together, we now summarize the four basic steps of a MB analysis:

1. Form a class  $\mathcal{B} = (\mu, \mathcal{S})$  of candidate beliefs, the choice of which may depend on (a) the assertions of interest, (b) the nature of your personal uncertainty, and/or (c) intuition and geometric/computational simplicity.
2. Choose the desired credibility level  $\alpha$ .
3. Employ a stochastic approximation algorithm to find an  $\alpha$ -MB as determined by the solution of (3.12).

4. Compute the posterior belief and plausibility functions via Monte Carlo integration by simulating the PRSs  $\mathcal{S}_{\omega(\alpha)}(U)$ .

In Sections 4 and 5 we will describe several specific classes of beliefs and the corresponding PRSs. These examples certainly will not exhaust all of the possibilities; they do, however, shed light on the considerations to be taken into account when constructing a class  $\mathcal{B}$  of beliefs.

#### 4. HIGH-DIMENSIONAL TESTING

A major focus of current statistical research is very-high-dimensional inference and, in particular, multiple testing. This is partly due to new scientific technologies, such as DNA microarrays and medical imaging devices, that give experimenters access to enormous amounts of data. A typical problem is to make inference on an unknown  $\Theta \in \mathbb{R}^n$  based on an observed  $X \sim N_n(\Theta, I_n)$ ; for example, testing  $H_{0i} : \Theta_i = 0$  for each  $i = 1, \dots, n$ . See Zhang and Liu [25] for a maximal belief solution of this many-normal-means problem. Below we consider a related problem—testing homogeneity of a Poisson process.

Suppose we monitor a system over a pre-specified interval of time, say,  $[0, \tau]$ . During that period of time, we observe  $n$  events/arrivals at times  $0 = \tau_0 < \tau_1 < \tau_2 < \dots < \tau_n$ , where the  $(n + 1)$ st event, taking place at  $\tau_{n+1} > \tau$ , is unobserved. Assume an exponential model for the inter-arrival times  $X_i = \tau_i - \tau_{i-1}$ ,  $i = 1, \dots, n$ , that is,

$$(4.1) \quad X_i \sim \text{Exp}(\Theta_i), \quad i = 1, \dots, n,$$

where the  $X_i$ 's are independent and the exponential rates  $\Theta_1, \dots, \Theta_n > 0$  are unknown. A question of interest is whether the underlying process is homogeneous, that is, whether the rates  $\Theta_1, \dots, \Theta_n$  have a common value. This question, or hypothesis, corresponds to the assertion

$$(4.2) \quad \begin{aligned} \mathcal{A} &= \{\text{the process is homogeneous}\} \\ &= \{\Theta_1 = \Theta_2 = \dots = \Theta_n\}. \end{aligned}$$

Let  $(X, \Theta)$  be the real-world quantities of interest, where  $X = (X_1, \dots, X_n)$ ,  $\Theta = (\Theta_1, \dots, \Theta_n)$  and  $\mathbb{X} = \mathbb{T} = (0, \infty)^n$ . Define the auxiliary variable  $U = (R, P)$ , where  $R > 0$  and  $P = (P_1, \dots, P_n)$  is in the  $(n - 1)$ -dimensional probability simplex  $\mathbb{P}_{n-1} \subset \mathbb{R}^n$ , defined as

$$\mathbb{P}_{n-1} = \left\{ (p_1, \dots, p_n) \in [0, 1]^n : \sum_{i=1}^n p_i = 1 \right\}.$$

The variables  $R$  and  $P$  are functions of the data  $X_1, \dots, X_n$  and the parameters  $\Theta_1, \dots, \Theta_n$ . The equation  $X = a(\Theta, U)$ , in this case, is given by  $X_i = RP_i/\Theta_i$ , where

$$(4.3) \quad R = \sum_{j=1}^n \Theta_j X_j \quad \text{and} \\ P_i = \frac{\Theta_i X_i}{\sum_{j=1}^n \Theta_j X_j}, \quad i = 1, \dots, n.$$

To complete the specification of the sampling model, we must choose the pivotal measure  $\mu$  for the auxiliary variable  $U = (R, P)$ . Given the nature of these variates, a natural choice is the product measure

$$(4.4) \quad \mu = \text{Gamma}(n, 1) \times \text{Unif}(\mathbb{P}_{n-1}).$$

The measure  $\mu$  in (4.4) is, indeed, consistent with the exponential model (4.1). To see this, note that  $\text{Unif}(\mathbb{P}_{n-1})$  is equivalent to the Dirichlet distribution  $\text{Dir}(1_n)$ , where  $1_n$  is an  $n$ -vector of unity. Then, conditional on  $(\Theta_1, \dots, \Theta_n)$ , it follows from standard properties of the Dirichlet distribution that  $\Theta_1 X_1, \dots, \Theta_n X_n$  are i.i.d.  $\text{Exp}(1)$ , which is equivalent to (4.1).

We now proceed with the WB analysis. Step 1 is to define the class of mappings  $\mathcal{S}$  for prediction of the unobserved auxiliary variables  $U^* = (R^*, P^*)$ . To expand a random draw  $U = (R, P) \sim \mu$  to a random set,

consider the class of maps  $\mathcal{S} = \{\mathcal{S}_\omega : \omega \in [0, \infty]\}$  defined as

$$(4.5) \quad \mathcal{S}_\omega(U) = \{(r, p) \in [0, \infty) \times \mathbb{P}_{n-1} : \\ K(P, p) \leq \omega\},$$

where  $K(P, p)$  is the Kullback–Leibler (KL) divergence

$$(4.6) \quad K(P, p) = \sum_{i=1}^n P_i \log(P_i/p_i), \quad p, P \in \mathbb{P}_{n-1}.$$

Several comments on the choice of PRSs (4.5) are in order. First, notice that  $\mathcal{S}_\omega(U)$  does not constrain the value of  $R$ , that is,  $\mathcal{S}_\omega(U)$  is just a cylinder in  $[0, \infty) \times \mathbb{P}_{n-1}$  defined by the  $P$ -component of  $U$ . Second, the use of the KL divergence in (4.5) is motivated by the correspondence between  $\mathbb{P}_{n-1}$  and the set of all probability measures on  $\{1, 2, \dots, n\}$ . The KL divergence is a convenient tool for defining neighborhoods in  $\mathbb{P}_{n-1}$ . Figure 4 shows cross-sections of several random sets  $\mathcal{S}_\omega(U)$  in the case of  $n = 3$ .

After choosing a credibility level  $\alpha \in (0, 1)$ , we are on to Step 3 of the analysis: finding an  $\alpha$ -MB. As in Section 3, define

$$Q_\omega(r, p) = \mu\{(R, P) : \mathcal{S}_\omega(R, P) \ni (r, p)\},$$

and, finally, choose  $\omega = \omega(\alpha)$  to solve the equation

$$\mu\{(R^*, P^*) : Q_\omega(R^*, P^*) \geq 1 - \alpha\} = \alpha.$$

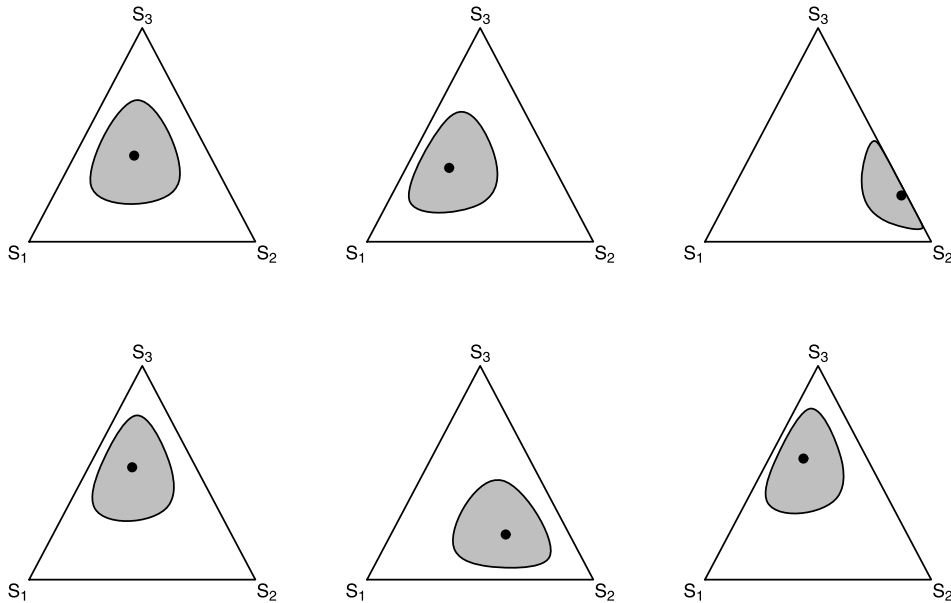


FIG. 4. Six realizations of  $R$ -cross sections of the PRS  $\mathcal{S}_\omega(R, P)$  in (4.5) in the case of  $n = 3$ . Here  $\mathbb{P}_2$  is the triangular region in the Barycentric coordinate system.

This calculation requires stochastic approximation.

For Step 4, first define the mapping  $\widehat{P}: \mathbb{T} \rightarrow \mathbb{P}_{n-1}$  by the component-wise formula  $\widehat{P}_i(\Theta) = \Theta_i X_i / \sum_j \Theta_j X_j$ ,  $i = 1, \dots, n$ . For inference on  $\Theta = (\Theta_1, \dots, \Theta_n)$ , a posterior focal element is of the form

$$M_X(R, P; \mathcal{S}_{\omega(\alpha)}) = \{\Theta : K(P, \widehat{P}(\Theta)) \leq \omega(\alpha)\}.$$

For the homogeneity assertion  $\mathcal{A}$  in (4.2) the posterior belief function is zero, but the plausibility is given by

$$\begin{aligned} \text{Pl}_X(\mathcal{A}; \mathcal{S}_{\omega(\alpha)}) \\ = 1 - \mu\{(R, P) : K(P, \widehat{P}(1_n)) > \omega(\alpha)\}, \end{aligned}$$

where  $\widehat{P}_i(1_n) = X_i / \sum_j X_j$ . Since  $\widehat{P}(1_n)$  is known and  $P \sim \text{Unif}(\mathbb{P}_{n-1})$  is easy to simulate, once  $\omega(\alpha)$  is available, the plausibility can be readily calculated using Monte Carlo.

In order to assess the performance of the MB method above in testing homogeneity, we will compare it with the typical likelihood ratio (LR) test. Let  $\ell(\Theta)$  be the likelihood function under the general model (4.1). Then the LR test statistic for  $H_0: \Theta_1 = \dots = \Theta_n$  is given by

$$\begin{aligned} L_0 &= \frac{\sup\{\ell(\Theta) : \Theta \in H_0\}}{\sup\{\ell(\Theta) : \Theta \in H_0 \cup H_0^c\}} \\ &= \left[ \frac{(\prod_{i=1}^n X_i)^{1/n}}{\bar{X}} \right]^n, \end{aligned}$$

a power of the ratio of the geometric and arithmetic means. If  $\widehat{P}$  is as defined before, then a little algebra shows that

$$L = -\log L_0 = nK(u_n, \widehat{P}(1_n)),$$

where  $u_n$  is the  $n$ -vector  $n^{-1}1_n$  which corresponds to the uniform distribution on  $\{1, 2, \dots, n\}$ . Note that

this problem is invariant under the group of scale transformations, so the null distribution of  $\widehat{P}(1_n)$  and, hence  $L$ , is independent of the common value of the rates  $\Theta_1, \dots, \Theta_n$ . In fact, under the homogeneity assertion (4.2),  $\widehat{P}(1_n) \sim \text{Unif}(\mathbb{P}_{n-1})$ .

EXAMPLE 6. To compare the MB and LR tests of homogeneity described above, we performed a simulation. Take  $n = n_1 + n_2 = 100$ ,  $n_1$  of the rates  $\Theta_1, \dots, \Theta_n$  to be 1 and  $n_2$  of the rates to be  $\theta$ , for various values of  $\theta$ . For each of 1000 simulated data sets, the plausibility for  $\mathcal{A}$  in (4.2). To perform the hypothesis test using  $q$ , we choose a nominal 5% level and say “reject the homogeneity hypothesis if plausibility  $< 0.05$ .” The power of the two tests are summarized in Figure 5, where we see that the MB test is noticeably better than the LR test. The MB test also controls the frequentist Type-I error at 0.05. But note that, unlike the LR test, the MB test is based on a meaningful data-dependent measure of the amount of evidence supporting the homogeneity assertion.

## 5. NONPARAMETRICS

A fundamental problem in nonparametric inference is the so-called *one-sample test*. Specifically, assume that  $X_1, \dots, X_n$  are i.i.d. observations from a distribution on  $\mathbb{R}$  with CDF  $F$  in a class  $\mathbb{F}$  of CDFs; the goal is to test  $H_0: F \in \mathbb{F}_0$  where  $\mathbb{F}_0 \subset \mathbb{F}$  is given. One application is a test for normality, that is, where  $\mathbb{F}_0 = \{N(\theta, \sigma^2) \text{ for some } \theta \text{ and } \sigma^2\}$ . This is an important problem, since many popular methods in applied statistics, such as regression and analysis of variance, often require an approximate normal distribution of the data, of residuals, etc.

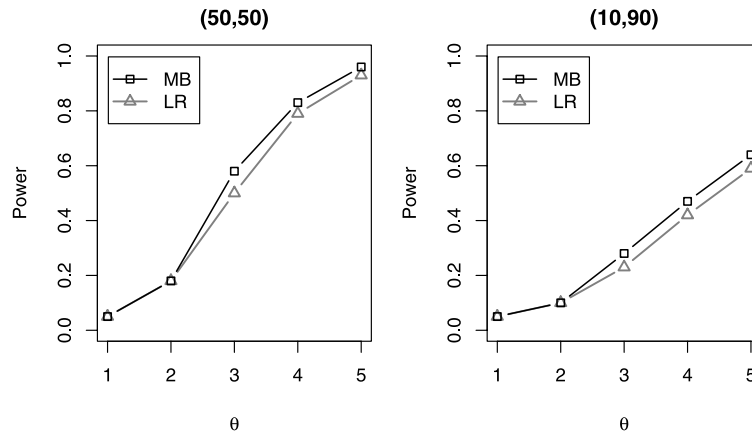


FIG. 5. Power of the MB and LR tests of homogeneity in Example 6, where  $\theta$  is the ratio of the rate for the last  $n_2$  observations to the rate of the first  $n_1$  observations. Left:  $(n_1, n_2) = (50, 50)$ . Right:  $(n_1, n_2) = (10, 90)$ .

We restrict attention to the simple one-sample testing problem, where  $\mathbb{F}_0 = \{F_0\} \subset \mathbb{F}$  is a singleton. Our starting point is the a-equation

$$(5.1) \quad X_i = F^{-1}(U_i), \quad F \in \mathbb{F}, i = 1, \dots, n,$$

where  $U_1, \dots, U_n$  are i.i.d.  $\text{Unif}(0, 1)$ . Since  $F$  is monotonically increasing, it is sufficient to consider the ordered data  $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ , the corresponding ordered auxiliary variables  $\tilde{U} = (U_{(1)}, \dots, U_{(n)})$ , and pivotal measure  $\mu$  determined by the distribution of  $\tilde{U}$ .

In this section we present a slightly different form of WB analysis based on *hierarchical* PRSs. In hierarchical Bayesian analysis, a random prior is taken to add an additional layer of flexibility. The intuition here is similar, but we defer the discussion and technical details to Appendix A.2.

For predicting  $\tilde{U}^*$ , we consider a class of beliefs indexed by  $\Omega = [0, \infty]$ , whose PRSs are small  $n$ -boxes inside the unit  $n$ -box  $[0, 1]^n$ . Start with a fixed set-valued mapping that takes ordered  $n$ -vectors  $\tilde{u} \in [0, 1]^n$ , points  $z \in (0.5, 1)$ , and forms the intervals  $[A_i(z), B_i(z)]$ , where

$$(5.2) \quad \begin{aligned} A_i(z) &= \text{qBeta}(p_i - zp_i \mid i, n + 1 - i), \\ B_i(z) &= \text{qBeta}(p_i + z(1 - p_i) \mid i, n + 1 - i) \end{aligned}$$

and  $p_i = \text{pBeta}(u_{(i)} \mid i, n - i + 1)$ . Here  $\text{pBeta}$  and  $\text{qBeta}$  denote CDF and inverse CDF of the Beta distribution, respectively. Then the mapping  $\mathcal{S}(\tilde{u}, z)$  is just the Cartesian product of these  $n$  intervals; cf. Example 5. Now sample  $\tilde{U}$  and  $Z$  from a suitable distribution depending on  $\omega$ :

- Take a draw  $\tilde{U}$  of  $n$  ordered  $\text{Unif}(0, 1)$  variables.
- Take  $V \sim \text{Beta}(\omega, 1)$  and set  $Z = \frac{1}{2}(1 + V)$ .

The result is a random set  $\mathcal{S}(\tilde{U}, Z) \in 2^{\mathbb{U}}$ . We call this approach “hierarchical” because one could first sample  $Z = z$  from the transformed beta distribution indexed by  $\omega$ , fix the map  $\mathcal{S}(\cdot, z)$ , and then sample  $\tilde{U}$ .

For a draw  $(\tilde{U}, Z)$ , the posterior focal elements for  $F$  look like

$$M_X(\tilde{U}, Z; \mathcal{S}) = \{F : A_i(Z) \leq F(X_{(i)}) \leq B_i(Z), \forall i = 1, \dots, n\}.$$

Details of the credibility of in a more general context are given in Appendix A.2. Stochastic approximation is used, as in Section 4, to optimize the choice of  $\omega$ . The MB method uses the posterior focal elements above, with optimal  $\omega$ , to compute the posterior belief and plausibility functions for the assertion  $\mathcal{A} = \{F = F_0\}$  of interest.

**EXAMPLE 7.** To illustrate the performance of the MB method, we present a small simulation study. We take  $F_0$  to be the CDF of a  $\text{Unif}(0, 1)$  distribution. Samples  $X_1, \dots, X_n$ , for various sample sizes  $n$ , are taken from several nonuniform distributions and the power of MB, along with some of the classical tests, is computed. We have chosen our nonuniform alternatives to be  $\text{Beta}(\beta_1, \beta_2)$  for various values of  $(\beta_1, \beta_2)$ . For the MB test, we use the decision rule “reject  $H_0$  if plausibility  $< 0.05$ .” Figure 6 shows the power of the level  $\alpha = 0.05$  Kolmogorov–Smirnov (KS), Anderson–Darling (AD), Cramér–von Mises (CV) and MB tests, as functions of the sample size  $n$  for six pairs of  $(\beta_1, \beta_2)$ . From the plots we see that the MB test outperforms the three classical tests in terms of power in all cases, in particular, when  $n$  is relatively small and the alternative is symmetric and “close” to the null [i.e., when  $(\beta_1, \beta_2) \approx (1, 1)$ ]. Here, as in Example 6, the MB test also controls the Type-I error at level  $\alpha = 0.05$ .

## 6. DISCUSSION

In this paper we have considered an modification of the DS theory in which some desired frequency properties can be realized while, at the same time, the essential components of DS inference, such as “don’t know,” remain intact. The WB method was justified within a more general framework of inferential models, where posterior probability-based inference with frequentist properties is the primary goal. In two high-dimensional hypothesis testing problems, the MB method performs quite well compared to popular frequentist methods in terms of power—more work is needed to fully understand this relationship between WB/MB hypothesis testing and frequentist power. Also, the detail in which these examples were presented should shed light on how MB can be applied in practice.

One potential criticism of the WB method is the lack of uniqueness of the a-equations and PRS mappings  $\mathcal{S}$ . At this stage, there are no optimality results justifying any particular choices. Our approach thus far has been to consider relatively simple and intuitive ways of constructing PRSs, but further research is needed to define these optimality criteria and to design PRSs that satisfy these criteria.

In addition to the applications shown above, preliminary results of WB methods in other statistical problems are quite promising. We hope that this work on WBs will inspire both applied and theoretical statisticians to take another look at what DS has to offer.

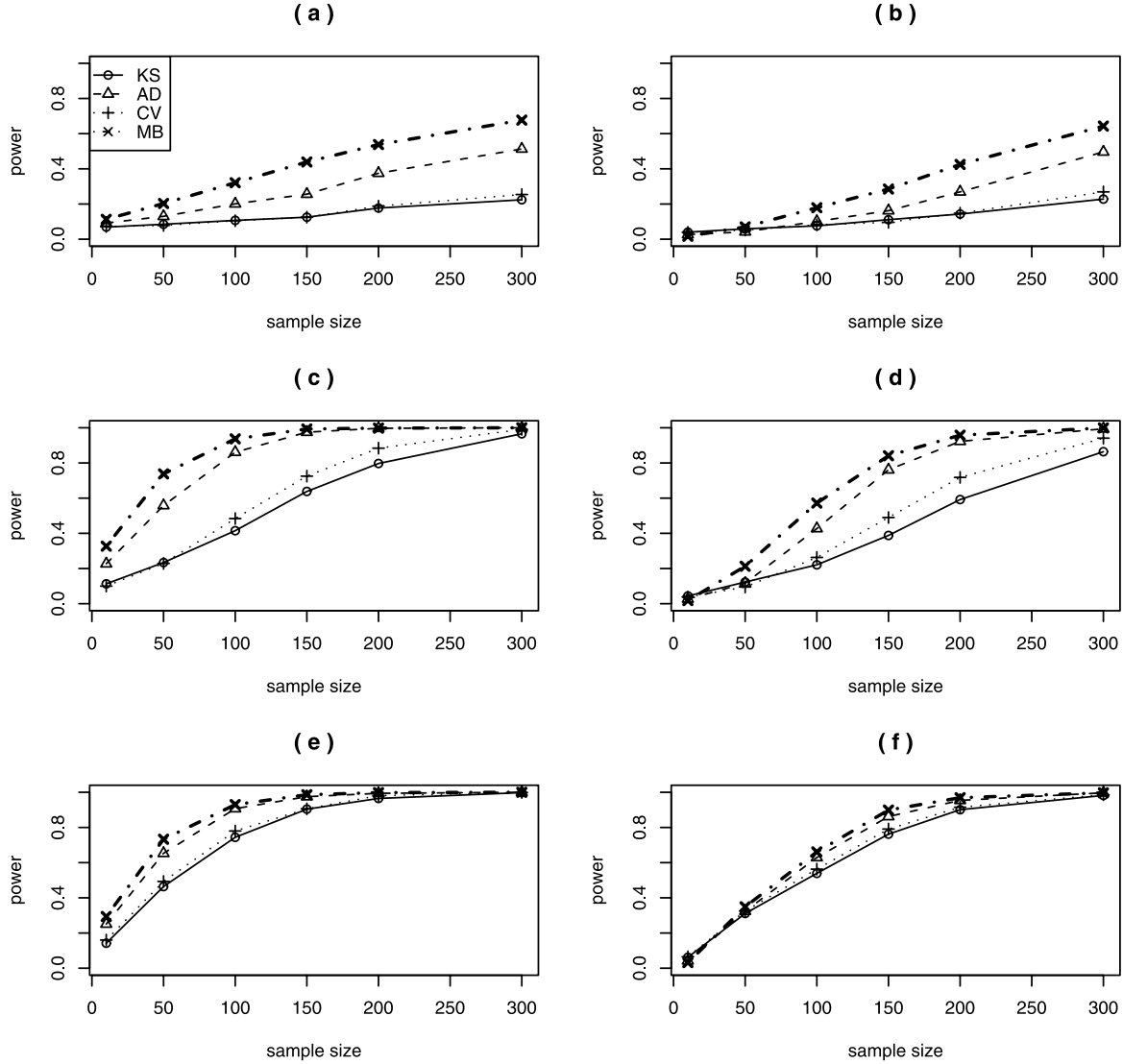


FIG. 6. Power comparison for the one-sample tests in Example 7 at level  $\alpha = 0.05$  for various values of  $n$ . The six alternatives are (a) Beta(0.8, 0.8); (b) Beta(1.3, 1.3); (c) Beta(0.6, 0.6); (d) Beta(1.6, 1.6); (e) Beta(0.6, 0.8); (f) Beta(1.3, 1.6).

## APPENDIX: TECHNICAL RESULTS

### A.1 Existence of a MB

Consider a class  $\mathcal{S} = \{\mathcal{S}_\omega : \omega \in \Omega\}$  of set-valued mappings. Assume that the index set  $\Omega$  is a complete metric space. Each  $\mathcal{S}_\omega$ , together with the pivotal measure  $\mu$ , define a belief function  $\mu_{\mathcal{S}_\omega^{-1}}$  on  $\mathbb{U}$ . Here we show that there is a  $\omega = \omega(\alpha)$  that solves the equation (3.12). To this end, we make the following assumptions:

- A1. Both the *conventional* and *vacuous* beliefs are encoded in  $\mathcal{S}$ .
- A2. If  $\omega_n \rightarrow \omega$ , then  $\mathcal{S}_{\omega_n}(u) \rightarrow \mathcal{S}_\omega(u)$  for each  $u \in \mathbb{U}$ .

Condition A1 is to make sure that  $\mathcal{B}$  is suitably rich, while A2 imposes a sort of continuity on the sets  $\mathcal{S}_\omega \in \mathcal{S}$ .

**PROPOSITION 1.** *Under assumptions A1–A2, there exists a solution  $\omega(\alpha)$  to (3.12) for any  $\alpha \in (0, 1)$ .*

**PROOF.** For notational simplicity, we write  $Q(\omega, u)$  for  $Q_\omega(u)$ . We start by showing  $Q(\omega, u)$  is continuous in  $\omega$ . Choose  $\omega \in \Omega$  and a sequence  $\omega_n \rightarrow \omega$ . Then under A2

$$\begin{aligned} Q(\omega_n, u) &= \int I_{\{\mathcal{S}_{\omega_n}(v) \not\ni u\}} d\mu(v) \\ &\rightarrow \int I_{\{\mathcal{S}_\omega(v) \not\ni u\}} d\mu(v) = Q(\omega, u) \end{aligned}$$

by the dominated convergence theorem (DCT). Since  $\omega_n \rightarrow \omega$  was arbitrary and  $\Omega$  is a metric space, it follows that  $Q(\cdot, u)$  is continuous on  $\Omega$ .

Write  $\varphi(\omega)$  for  $\varphi_\alpha(\omega)$  in (3.9); we will now show that  $\varphi(\cdot)$  is continuous. Again choose  $\omega \in \Omega$  and a sequence  $\omega_n \rightarrow \omega$ . Define  $J_\omega(u) = I_{\{Q(\omega, u) \geq 1 - \alpha\}}$ , so that  $\varphi(\omega) = \int J_\omega(u) d\mu(u)$ . Since

$$|\varphi(\omega_n) - \varphi(\omega)| \leq \int |J_{\omega_n}(u) - J_\omega(u)| d\mu(u)$$

and the integrand on the right-hand side is bounded by 2, it follows, again follows by the DCT, that  $\varphi(\omega_n) \rightarrow \varphi(\omega)$  and, hence, that  $\varphi(\cdot)$  is continuous on  $\Omega$ . But A1 implies that  $\varphi(\cdot)$  takes values 0 and 1 on  $\Omega$  so by the intermediate value theorem, for any  $\alpha \in (0, 1)$ , there exists a solution  $\omega = \omega(\alpha)$  to the equation  $\varphi(\omega) = \alpha$ .  $\square$

## A.2 Hierarchical PRSs

In Section 5 we considered a WB analysis with hierarchical PRSs. The purpose of this generalization is to provide a more flexible choice of random sets for predicting the unobserved  $U^*$ . Here we give a theoretical justification along the lines in Section 3.4.

Let  $\omega \in \Omega$  index a family of probability measures  $\lambda_\omega$  on a space  $\mathbb{Z}$ , and suppose  $\mathcal{S}(\cdot, \cdot)$  is a fixed set-valued mapping  $\mathbb{U} \times \mathbb{Z} \rightarrow 2^{\mathbb{U}}$ , assumed to satisfy  $U \in \mathcal{S}(U, Z)$  for all  $Z$ . A hierarchical PRS is defined by first taking  $Z \sim \lambda_\omega$  and then choosing the map  $\mathcal{S}_Z(\cdot) = \mathcal{S}(\cdot, Z)$  defined on  $\mathbb{U}$ . This amounts to a product pivotal measure  $\mu \times \lambda_\omega$ . Toward credibility of  $(\mu \times \lambda_\omega, \mathcal{S})$ , define the noncoverage probability

$$\begin{aligned} \overline{Q}_\omega(u) &= (\mu \times \lambda_\omega)\{(U, Z) : \mathcal{S}(U, Z) \not\ni u\} \\ &= \int Q_z(u) d\lambda_\omega(z), \end{aligned}$$

a mixture of the noncoverage probabilities in (3.8). Then we have the following, more general, definition of credibility.

**DEFINITION 4.**  $(\mu, \mathcal{S}_\omega)$  is credible at level  $\alpha$  if

$$\overline{\varphi}_\alpha(\omega) := \mu\{U^* : \overline{Q}_\omega(U^*) \geq 1 - \alpha\} \leq \alpha.$$

Beliefs which are credible in the sense of Definition 1 are also credible according to Definition 4—take  $\lambda_\omega$  to be a point mass at  $\omega$ . It is also clear that if  $(\mu, \mathcal{S}_z)$  is credible in the sense of Definition 1 for all  $z \in \mathbb{Z}$ , then  $(\mu \times \lambda_\omega, \mathcal{S})$  will also be credible. Next we generalize Theorem 1 to handle the case of hierarchical PRSs.

**THEOREM 2.** Suppose that  $(\mu \times \lambda_\omega, \mathcal{S})$  is credible at level  $\alpha$  in the sense of Definition 4, and that  $(\mu \times \lambda_\omega)\{(U, Z) : M_X(U, Z; \mathcal{S}) \neq \emptyset\} = 1$ . Then for any assertion  $\mathcal{A} \subset \mathbb{T}$ , the belief function  $\text{Bel}_X(\mathcal{A}; \mathcal{S}) = (\mu \times \lambda_\omega)\mathcal{S}^{-1}(\mathcal{A})$  satisfies

$$\text{P}_\Theta\{\text{Bel}_X(\mathcal{A}; \mathcal{S}) \geq 1 - \alpha\} \leq \alpha, \quad \Theta \in \mathcal{A}^c.$$

**PROOF.** Start by fixing  $Z = z$ , and write  $\mathcal{S}_z(\cdot) = \mathcal{S}(\cdot, z)$ . For  $\Theta \in \mathcal{A}^c$ , monotonicity of the belief function gives

$$\begin{aligned} \text{Bel}_X(\mathcal{A}; \mathcal{S}_z) &\leq \text{Bel}_X(\{\Theta\}^c; \mathcal{S}_z) \\ &= \mu\{U : M_X(U; \mathcal{S}_z) \not\ni \Theta\}. \end{aligned}$$

When  $\Theta$  is the true value, the event  $M_X(U; \mathcal{S}_z) \not\ni \Theta$  is equivalent to  $\mathcal{S}_z(U) \not\ni U^*$ ; consequently,

$$\text{Bel}_X(\mathcal{A}; \mathcal{S}_z) \leq \mu\{U : \mathcal{S}_z(U) \not\ni U^*\} = Q_z(U^*).$$

For the hierarchical PRS, the belief function satisfies

$$\begin{aligned} \text{Bel}_X(\mathcal{A}; \mathcal{S}) &= (\mu \times \lambda_\omega)\{(U, Z) : M_X(U, Z; \mathcal{S}) \subseteq \mathcal{A}\} \\ &= \int \mu\{U : M_X(U; \mathcal{S}_z) \subseteq \mathcal{A}\} d\lambda_\omega(z) \\ &= \int \text{Bel}_X(\mathcal{A}; \mathcal{S}_z) d\lambda_\omega(z) \\ &\leq \int Q_z(U^*) d\lambda_\omega(z) \\ &= \overline{Q}_\omega(U^*). \end{aligned}$$

The claim now follows from credibility of the belief  $(\mu \times \lambda_\omega, \mathcal{S})$ .  $\square$

## ACKNOWLEDGMENTS

The authors would like to thank Professor A. P. Dempster for sharing his insight, and also the Editor, Associate Editor and three referees for helpful suggestions and criticisms.

## REFERENCES

- [1] DEMPSTER, A. P. (1963). Further examples of inconsistencies in the fiducial argument. *Ann. Math. Statist.* **34** 884–891. [MR0150865](#)
- [2] DEMPSTER, A. P. (1966). New methods for reasoning towards posterior distributions based on sample data. *Ann. Math. Statist.* **37** 355–374. [MR0187357](#)
- [3] DEMPSTER, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.* **38** 325–339. [MR0207001](#)
- [4] DEMPSTER, A. P. (1968). A generalization of Bayesian inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **30** 205–247. [MR0238428](#)

- [5] DEMPSTER, A. P. (1969). Upper and lower probability inferences for families of hypotheses with monotone density ratios. *Ann. Math. Statist.* **40** 953–969. [MR0246427](#)
- [6] DEMPSTER, A. P. (2008). Dempster–Shafer calculus for statisticians. *Internat. J. Approx. Reason.* **48** 265–277. [MR2419025](#)
- [7] DENOEU, T. (2006). Constructing belief functions from sample data using multinomial confidence regions. *Internat. J. Approx. Reason.* **42** 228–252. [MR2246874](#)
- [8] EDLEFSEN, P. T., LIU, C. and DEMPSTER, A. P. (2009). Estimating limits from Poisson counting data using Dempster–Shafer analysis. *Ann. Appl. Statist.* **3** 764–790.
- [9] FISHER, R. A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society* **26** 528–535.
- [10] FISHER, R. A. (1935). The logic of inductive inference. *J. Roy. Statist. Soc.* **98** 39–82.
- [11] FRASER, D. A. S. (1968). *The Structure of Inference*. Wiley, New York. [MR0235643](#)
- [12] HANNIG, J. (2009). On generalized fiducial inference. *Statist. Sinica* **19** 491–544. [MR2514173](#)
- [13] KOHLAS, J. and MONNEY, P.-A. (2008). An algebraic theory for statistical information based on the theory of hints. *Internat. J. Approx. Reason.* **48** 378–398. [MR2419014](#)
- [14] KUSHNER, H. J. and YIN, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed. Springer, New York. [MR1993642](#)
- [15] LINDLEY, D. V. (1958). Fiducial distributions and Bayes’ theorem. *J. Roy. Statist. Soc. Ser. B* **20** 102–107. [MR0095550](#)
- [16] MARTIN, R. and GHOSH, J. K. (2008). Stochastic approximation and Newton’s estimate of a mixing distribution. *Statist. Sci.* **23** 365–382. [MR2483909](#)
- [17] ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Statist.* **22** 400–407. [MR0042668](#)
- [18] SHAFER, G. (1976). *A Mathematical Theory of Evidence*. Princeton Univ. Press, Princeton, NJ. [MR0464340](#)
- [19] SHAFER, G. (1978/79). Nonadditive probabilities in the work of Bernoulli and Lambert. *Arch. Hist. Exact Sci.* **19** 309–370. [MR0515919](#)
- [20] SHAFER, G. (1979). Allocations of probability. *Ann. Probab.* **7** 827–839. [MR0542132](#)
- [21] SHAFER, G. (1981). Constructive probability. *Synthese* **48** 1–60. [MR0623413](#)
- [22] SHAFER, G. (1982). Belief functions and parametric models (with discussion). *J. Roy. Statist. Soc. Ser. B* **44** 322–352. [MR0693232](#)
- [23] YAGER, R. and LIU, L. (eds.) (2008). *Classic Works of the Dempster–Shafer Theory of Belief Functions. Stud. Fuzziness Soft Comput.* **219**. Springer, Berlin. [MR2458525](#)
- [24] ZABELL, S. L. (1992). R. A. Fisher and the fiducial argument. *Statist. Sci.* **7** 369–387. [MR1181418](#)
- [25] ZHANG, J. and LIU, C. (2010). Dempster–Shafer inference with weak beliefs. *Statistica Sinica*. To appear.