General Theory of Inferential Models I.
Conditional Inference

by

Ryan Martin
Indiana University-Purdue University Indianapolis

Jing-Shiang Hwang
Academica Sinica

Chuanhai Liu
Purdue University

Technical Report #10-04

Department of Statistics
Purdue University

November 2010

◇

GENERAL THEORY OF INFERENTIAL MODELS I. CONDITIONAL INFERENCE

By Ryan Martin, Jing-Shiang Hwang, and Chuanhai Liu

Indiana University-Purdue University Indianapolis, Academia Sinica, and Purdue University

As applied problems have grown more complex, statisticians have been gradually led to reconsider the foundations of statistical inference. The recently proposed inferential model (IM) framework of Martin, Zhang and Liu (2010) achieves an interesting compromise between the Bayesian and frequentist ideals. Indeed, inference is based on posterior probability-like quantities, but there are no priors and the inferential output satisfies certain desirable long-run frequency properties. In this two-part series, we further develop the theory of IMs into a general framework for statistical inference.

Here, in Part I, we build on the idea of making inference by predicting unobserved auxiliary variables, focusing primarily on an intermediate step of *conditioning*, whereby the dimension of this auxiliary variable is reduced to a manageable size. This dimension reduction step leads to a simpler construction of IMs having the required long-run frequency properties. We show that under suitable conditions, this dimension reduction step can be made without loss of information, and that these conditions are satisfied in a wide class of models, including those with a group invariance structure. The important credibility theorem of Zhang and Liu (2010) is extended to handle the case of conditional IMs, and connections to conditional inference in the likelihood framework are made which, in turn, allow for numerical approximation of the conditional posterior belief functions using the well-known "magic formula" of Barndorff-Nielsen (1983). The conditional IM approach is illustrated on a variety of examples, including Fisher's problem of the Nile.

1. Introduction. R. A. Fisher's brand of statistical inference (Fisher 1973) is often viewed as some sort of middle-ground between the purely Bayesian and purely frequentist approaches. Two important examples are his fiducial argument (Fisher 1935a; Zabell 1992) and his ideas on conditional inference (Fisher 1925, 1934, 1935b). As modern statistical problems have grown more complex, contemporary statisticians have been gradually led to reconsider the foundations of statistical inference. Perhaps influenced

AMS 2000 subject classifications: Primary 62A01; secondary 68T37, 62B05

Keywords and phrases: Ancillarity, Bayes, belief functions, conditional inference, frequentist, sufficiency, predictive random sets

by ideas of Fisher, a current focus is on achieving some sort of compromise between the Bayesian and frequentist ideals. Case in point is the recent work on objective Bayesian analysis with default/reference priors (Berger 2006; Berger, Bernardo and Sun 2009; Ghosh, Delampady and Samanta 2006). An important goal of objective Bayes analysis is to construct priors for which certain posterior inferences, such as credible intervals, closely match that of a frequentist, so the spirit of compromise is clear; the *calibrated* Bayesian analysis of Rubin (1984), Dawid (1985), and Little (2010) has similar motivations. But difficulties remain in choosing good reference priors for high-dimensional problems so, despite these efforts, a fully satisfactory objective Bayes theory has yet to emerge. Efron (1998) predicts that (i) statistical problems in the 21st century will require new tools with good Bayesian/frequentist properties, and (ii) that "something like fiducial inference" will be influential to the development of these tools. He was correct on the first part of his prediction: empirical Bayes methods, for example, have been influential in analyzing data coming from new high-throughput devices such as microarrays (Efron 2008, 2010). The second part of Efron's prediction remains uncertain. The goal of this series of three papers is to further develop the notion of *inferential models*—a new "something like fiducial inference"—into a general framework for statistical inference, partially validating the second part of Efron's prediction.

Recently, Martin, Zhang and Liu (2010) present a modification of the Dempster-Shafer theory of belief functions that has desirable Bayesian- and frequentist-like qualities in the statistical inference problem. Dempster-Shafer (DS) theory (Dempster 1966, 1967, 1968, 2008; Shafer 1976), an extension of Fisher's fiducial argument, gives a recipe for constructing a posterior belief function over the parameter space Θ using only the sampling model and observed data; in particular, no prior distribution over Θ is assumed or required. Although DS theory is widely used in computer science and engineering (Yager and Liu 2008), posterior belief functions for statistical inference have yet to be widely accepted, perhaps because the numerical values assigned by these conventional belief functions do not, in general, have the long-run frequency properties statisticians are familiar with. Martin, Zhang and Liu (2010) extend the ideas of Zhang and Liu (2010) and propose a framework of *inferential models* (IMs) in which the conventional posterior belief functions are suitably weakened so that certain long-run frequency properties are realized. These details may be unfamiliar to some readers (see Section 2), but that the resulting inferential method achieves a balance between Bayesian and frequentist ideals should be clear.

The empirical results in Zhang and Liu (2010) and Martin, Zhang and

Liu (2010) indicate the potential of this IM framework for a wide range of statistical problems—even high-dimensional problems—but the ideas and methods presented therein are somewhat ad hoc. In particular, only very loose guidelines for constructing IMs are given. In this two-part series of papers, we attempt to sharpen the ideas of Zhang and Liu (2010) and Martin, Zhang and Liu (2010) to form a general theory of IMs.

Here, in Part I, we stick with the basic but fundamental idea of making probabilistic statements about the parameter of interest by predicting an unobservable auxiliary variable; see Section 2 for a review. Our focus, however, is on a preliminary step of reducing the dimension of this auxiliary variable before attempting its prediction. We consider a dimension reduction step based on *conditioning*, and develop a framework of conditional IMs for making inference (hypothesis tests, confidence intervals, etc) on the unknown parameter. Important connections between our framework and Fisher's ideas of sufficiency and conditional inference are made in Section 3.5. After introducing the general conditional approach in Section 3.2, we use several relatively simple textbook examples in Section 3.3 to illustrate the main ideas. In Section 3.4 we extend the main theorem of Martin, Zhang and Liu (2010), showing that the desirable long-run frequency properties are attained over a "relevant subset" of the sample space. Theoretical properties of conditional IMs are investigated in Section 4 for a broad class of sampling models which are invariant under a general group of transformation, and Section 5 provides details of a conditional IM analysis for Fisher's problem of the Nile. Some concluding remarks are made in Section 6.

2. Sampling model and statistical a-inference. In this section we describe the basic sampling model as well as review belief functions and the construction of inferential models (IMs). Some relatively simple examples will be considered in Section 2.4.

2.1. *Sampling model and a-equation.* The sampling model P_θ is a probability measure on $\mathbb{X}$ that encodes the joint distribution of the data vector $X = (X_1, \ldots, X_n)'$. As in Martin, Zhang and Liu (2010), we further assume that the sampling model P_θ is defined as follows. Take $\mathbb{U}$ to be a more-or-less arbitrary auxiliary space equipped with a probability measure ν. In what follows, we will occasionally attach the "a-" prefix to quantities/concepts related to this auxiliary space; that is, we call $\mathbb{U}$ the a-space, ν the a-measure, and so on. Let $a : \mathbb{U} \times \Theta \to \mathbb{X}$ be a measurable function. Now take a random draw $U \sim \nu$, and choose X such that

$$X = a(U, \theta). \tag{2.1}$$

In other words, the sampling model for X given θ, is determined by the a-measure ν on $\mathbb{U}$ and the constraints (2.1). More formally, if we write $a_\theta(u) = a(u, \theta)$ for fixed θ, then the sampling model P_θ is the push-forward measure $\mathsf{P}_\theta = \nu\, a_\theta^{-1}$ induced by the a-measure and the mapping a_θ. Martin, Zhang and Liu (2010) call (2.1) the *a-equation*. This a-equation and its extension (see Part II) will be of critical importance to our development.

2.2. *Belief functions.* The primary tools to be used for a-inference are *belief functions*, a generalization of probability measures, first introduced by Dempster (1967) and later formalized by Shafer (1976, 1979). The key property that distinguishes belief functions from probability measures is *subadditivity*: if $\text{Bel} : \mathscr{A} \to [0, 1]$ is a belief function defined on a collection $\mathscr{A}$ of measurable subsets of Θ, then

$$\text{Bel}(\mathcal{A}) + \text{Bel}(\mathcal{A}^c) \leq 1 \quad \text{for all } \mathcal{A} \in \mathscr{A}.$$

For probability measures, equality obtains for all $\mathcal{A}$, but not necessarily for belief functions. The intuition is that evidence that does not support $\mathcal{A}$ may not support $\mathcal{A}^c$ either. A related quantity is the *plausibility function*, defined as

$$\text{Pl}(\mathcal{A}) = 1 - \text{Bel}(\mathcal{A}^c).$$

It follows immediately from the subadditivity property of Bel that

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

For this reason, Bel and Pl have often been referred to as *lower* and *upper* probabilities, respectively (Dempster 1967). The plausibility will be particularly useful for designing statistical procedures based on a posterior belief function output; see Section 2.5.

Following the classical Dempster-Shafer theory for statistical inference, Martin, Zhang and Liu (2010) construct a basic belief function Bel_x on Θ as follows. For the general a-equation (2.1), define the set of data-parameter pairs which are consistent with the a-equation and a particular u in the a-space $\mathbb{U}$; that is,

$$M_x(u) = \{\theta \in \Theta : x = a(u, \theta)\}, \quad u \in \mathbb{U}.$$

Following Shafer (1976), we call $M_x(u)$ a *focal element* indexed by $u \in \mathbb{U}$. In general, the focal elements could be non-singletons, so $M_x(U)$ can be viewed as a random set when $U \sim \nu$. The belief function is then just the

(conditional) ν-probability that the random set $M_x(U)$, as a function of U, falls completely inside a particular set $\mathcal{A} \subset \Theta$. Specifically,

$$\text{Bel}_x(\mathcal{A}) = \frac{\nu\{u : M_x(u) \subseteq \mathcal{A},\ M_x(u) \neq \emptyset\}}{\nu\{u : M_x(u) \neq \emptyset\}}. \tag{2.2}$$

In all the cases we consider, the denominator will be equal to 1. That Bel_x in (2.2) has the subadditivity property in general is easy to see: $\text{Bel}_x(\mathcal{A})$ and $\text{Bel}_x(\mathcal{A}^c)$ tally the conditional ν-probability that $M_x(U)$ is completely inside and outside of $\mathcal{A}$, respectively, given $M_x(U) \neq \emptyset$, but in general there is positive ν-probability that $M_x(U)$ covers parts of both $\mathcal{A}$ and $\mathcal{A}^c$.

The interpretation of $\text{Bel}_x(\mathcal{A})$ is similar to that of a Bayesian posterior probability. Indeed, the belief function $\text{Bel}_x(\mathcal{A})$ encodes one's certainty that the true θ lies in $\mathcal{A}$ (Shafer 1976) based solely on the sampling model and data—no prior on Θ is necessary. But when the sampling model has a group structure, the belief function will often be the same as the fiducial and objective Bayes posterior distributions. Therefore, in these as well as other cases, the basic belief function will not be well-calibrated in the sense of Rubin (1984). IMs, reviewed in the next section, provide a general framework in which the basic belief function Bel_x may be suitably calibrated.

2.3. *Inferential models.* Martin, Zhang and Liu (2010) introduced the concept of inferential models (IMs) to extend the scope of posterior belief functions for statistical inference. The goal is to shrink the numerical values of the conventional belief function, but just enough so that certain desirable long-run frequency properties are realized.

The formal definition of an IM is quite simple. Let Bel_x be the basic belief function on Θ obtained in the previous section. An IM specifies a new or weaker belief function, say $\text{Bel}_x^\star$, on Θ such that

$$\text{Bel}_x^\star(\mathcal{A}) \leq \text{Bel}_x(\mathcal{A}) \quad \text{for all } \mathcal{A}. \tag{2.3}$$

That is, the new belief function is just a shrunken version of the basic belief function Bel_x. But one must control the amount by which Bel_x is shrunk in order to realize the desirable long-run frequency properties. The shrinking procedure used by Martin, Zhang and Liu (2010) and Zhang and Liu (2010) is called the method of *weak beliefs* and relies on what are called *predictive random sets* (PRSs).

Martin, Zhang and Liu (2010) and Zhang and Liu (2010) argue that inference on θ is equivalent to predicting the *true* but unobserved value $U^\star$ of the a-variable that corresponds to the actual observed data $X = x$ and the true value of θ. That is, this unobserved $U^\star$ must satisfy

$$x = a(U^\star, \theta) \quad \text{for the true } \theta.$$

Both fiducial and DS inference can be understood from this viewpoint, but they each try to predict $U^\star$ with a draw $U \sim \nu$. But, intuitively, this is overly optimistic since U will, in general, miss $U^\star$ with ν-probability 1. The method of weak beliefs relaxes this assumption and instead tries to hit the target $U^\star$ with a random set $\mathcal{S}(U) \supseteq \{U\}$, where $U \sim \nu$. The set $\mathcal{S}(U)$ is called a predictive random set (PRS). Some simple examples of PRSs can be found in Section 2.4.

For a given set-valued map $\mathcal{S}$, define the enlarged focal elements

$$M_x(u;\mathcal{S}) = \bigcup_{u' \in \mathcal{S}(u)} \{\theta : x = a(u', \theta)\}, \quad u \in \mathbb{U}. \tag{2.4}$$

It is obvious that $M_x(u;\mathcal{S}) \supseteq M_x(u)$, with equality if and only if $\mathcal{S}(u) = \{u\}$. Moreover, Martin, Zhang and Liu (2010) show that

$$\text{Bel}_x(\mathcal{A};\mathcal{S}) := \frac{\nu\{u : M_x(u;\mathcal{S}) \subseteq \mathcal{A},\ M_x(u;\mathcal{S}) \neq \emptyset\}}{\nu\{u : M_x(u;\mathcal{S}) \neq \emptyset\}} \tag{2.5}$$

defines a bonafide IM in the sense of (2.3). Note that taking $\mathcal{S}(U) = \{U\}$ produces the basic belief function (2.2). This trivial choice of PRS is not satisfactory for inference (Definition 1) but it will be used for comparing different a-equation/a-measure pairs (Definition 2).

It is intuitively clear that not every $\mathcal{S}$ will lead to good frequency properties of the belief function $\text{Bel}_X(\cdot;\mathcal{S})$. For this we must place some extra conditions on $\mathcal{S}$ in the form of coverage probabilities. Define

$$Q(u;\mathcal{S}) = \nu\{U : \mathcal{S}(U) \not\ni u\}, \quad u \in \mathbb{U}, \tag{2.6}$$

which is the probability that the PRS $\mathcal{S}(U)$ misses its target u. The condition imposed on $\mathcal{S}$ is that $Q(U^\star;\mathcal{S})$, a function of the unobserved $U^\star \sim \nu$, should be probabilistically small.

DEFINITION 1. $\mathcal{S} = \mathcal{S}(U)$ is *credible* for predicting $U^\star$ at level α if

$$\nu\{U^\star : Q(U^\star;\mathcal{S}) \geq 1-\alpha\} \leq \alpha. \tag{2.7}$$

The following theorem relates credibility of PRSs to desirable frequency properties of the enlarged belief function (2.5).

THEOREM 1 (Martin-Zhang-Liu). *Suppose $\mathcal{S}$ is credible for predicting $U^\star$ at level α, and $M_x(U;\mathcal{S}) \neq \emptyset$ with ν-probability 1 for all x. Then for any assertion $\mathcal{A} \subset \Theta$, $\text{Bel}_X(\mathcal{A};\mathcal{S})$ in (2.5), as a function of X, satisfies*

$$\mathsf{P}_\theta\{\text{Bel}_X(\mathcal{A};\mathcal{S}) \geq 1-\alpha\} \leq \alpha, \quad \forall\, \theta \in \mathcal{A}^c. \tag{2.8}$$

Therefore, IMs formed via the method of weak beliefs with credible PRSs will have the desired long-run frequency properties. The next section gives some simple examples, and Section 2.5 will show how Theorem 1 relates to the inference problem. We refer to Ermini Leaf and Liu (2010) for discussion on the case when the fundamental condition "$M_x(U;\mathcal{S}) \neq \emptyset$ *with* ν*-probability 1 for all* x" does not hold.

At this point we should emphasize that the IM approach is substantially different from the DS and fiducial theories. While belief functions are used to make inference, we do not adopt the fundamental operation of DS, namely Dempster's rule of combination (Shafer 1976, Ch. 3), and therefore we avoid the potential difficulties it entails (Ermini Leaf, Hui and Liu 2009; Walley 1987). Moreover, since the belief functions we use for inference are not probability measures, and we do not operate on them as if they were, our conclusions cannot coincide with those resulting from a fiducial (or Bayesian) analysis. There are similarities, however, to the imprecise prior Bayes approach (Walley 1996) and the robust Bayes approach (Berger 1984). But perhaps the most important difference is our focus on long-run frequency properties—DS and fiducial are void of such frequentist concerns.

2.4. *Examples.* Next we consider several relatively simple textbook-style examples; more complex models appear in Sections 3 and 5. For each example, the a-measure ν is just $\mathsf{Unif}(0,1)$, and for PRSs we will take the set-valued mapping

$$\mathcal{S}(u) = [u/2, (1+u)/2].$$

Other choices for $\mathcal{S}(u)$ include $[0,u]$, $[u,1]$, and $\{u' : |u'-0.5| \leq |u-0.5|\}$. These will produce slightly different IMs for the parameter of interest, but each can be shown to satisfy the conditions of Theorem 1.

EXAMPLE 1 (Normal model). Suppose X is a single sample from a $\mathsf{N}(\mu,\sigma^2)$ population. Assume, for simplicity, that the variance is known, say $\sigma^2 = 1$. We have the simple model $X = \mu + \Phi^{-1}(U)$, where Φ is the distribution function of $\mathsf{N}(0,1)$. The focal elements are singletons, so the basic belief function is a probability measure, namely the $\mathsf{N}(x,1)$ distribution. Example 4 of Martin, Zhang and Liu (2010) gives a general formula, along with plots, of a weakened belief function for a flexible class of PRSs.

EXAMPLE 2 (Exponential model). Suppose X is a single sample from a $\mathsf{Exp}(\lambda)$ population, where $\lambda > 0$ is a scale parameter. We have the simple model $X = \lambda F^{-1}(U)$, where F is the distribution function of $\mathsf{Exp}(1)$. For $\mathcal{A} = \{\lambda \leq \lambda_0\}$, the basic belief function is easily found to be

$$\mathrm{Bel}_x(\mathcal{A}) = \nu\{u : x \leq \lambda_0 F^{-1}(u)\} = \exp\{-x/\lambda_0\}.$$

For the choice of PRS $\mathcal{S}(u)$ above, and the same assertion $\mathcal{A} = \{\lambda \leq \lambda_0\}$, the weakened belief function is

$$\text{Bel}_x(\mathcal{A}; \mathcal{S}) = \nu\{u : x \leq \lambda_0 F^{-1}(u/2)\} = 1 - 2(1 - \exp\{-x/\lambda_0\}).$$

That $\text{Bel}_x(\mathcal{A}; \mathcal{S}) < \text{Bel}_x(\mathcal{A})$ is immediately clear.

The reader may notice that Examples 1 and 2 have a location-scale structure, and that the basic belief function (as well as the fiducial distribution) corresponds to the Bayesian posterior probability when the parameter is given the right-invariant Haar prior. It turns out that this is a general phenomenon in group invariant problems; see Fraser (1961).

EXAMPLE 3 (Bernoulli model). Flip a coin with probability θ of landing heads, and let X be 1 or 0 depending on whether the coin lands heads or tails. Then $X \sim \mathsf{Ber}(\theta)$. A simple a-equation for this model is $X = I_{\{U \leq \theta\}}$, where I_A is the indicator that event A occurs. The focal element is

$$M_x(u) = \begin{cases} [0, u] & \text{if } x = 0 \\ [u, 1] & \text{if } x = 1. \end{cases}$$

Notice that these focal elements are not singletons; therefore, the belief function cannot be a probability measure. For an assertion $\mathcal{A} = \{\theta \leq \theta_0\}$, the basic belief function is

$$\text{Bel}_x(\mathcal{A}) = \begin{cases} \theta_0 & \text{if } x = 0 \\ I_{\{\theta_0 = 1\}} & \text{if } x = 1. \end{cases}$$

The intuition here is that if tails is observed, then we believe that all values of θ are equally likely, but if heads is observed, then we cannot put any nontrivial upper bound on θ. For the PRS $\mathcal{S}(u)$ above, the new focal elements are easily seen to be

$$M_x(u; \mathcal{S}) = \begin{cases} [0, (1+u)/2] & \text{if } x = 0 \\ [u/2, 1] & \text{if } x = 1, \end{cases}$$

so the weakened belief function is

$$\text{Bel}_x(\mathcal{A}) = \begin{cases} \max\{2\theta_0 - 1, 0\} & \text{if } x = 0 \\ I_{\{\theta_0 = 1\}} & \text{if } x = 1. \end{cases}$$

We should point out that this IM is a bit conservative for the particular assertion $\mathcal{A}$ in question since the basic focal elements $M_x(u)$ themselves are not singletons.

EXAMPLE 4 (Poisson model). In this example we find that the a-equation need not have a nice simple expression; here, as well as in other discrete problems, the a-equation is just a rule for defining data based on a given parameter and a-variable. Suppose X is a single sample from a $\mathsf{Poi}(\lambda)$ population. Then the sampling model can be written as

$$F_\lambda(X-1) \leq U < F_\lambda(X), \quad U \sim \mathsf{Unif}(0,1),$$

where F_λ is the distribution function of $\mathsf{Poi}(\lambda)$. Integration-by-parts reveals that $F_\lambda(x) = 1 - G_{x+1}(\lambda)$, where $G_\alpha(\lambda)$ is a $\mathsf{Gam}(\alpha, 1)$ distribution function. Then the sampling model above is equivalent to

$$G_{X+1}(\lambda) \leq U < G_X(\lambda), \quad U \sim \mathsf{Unif}(0,1),$$

where we have used the fact that U and $1-U$ are both $\mathsf{Unif}(0,1)$. Therefore, given $X = x$, the basic focal elements are intervals of the form

$$M_x(u) = \left[G_{x+1}^{-1}(u), G_x^{-1}(u) \right).$$

For an assertion $\mathcal{A} = \{\lambda \leq \lambda_0\}$, the basic belief function is

$$\text{Bel}_x(\mathcal{A}) = \nu\{u : G_x^{-1}(u) \leq \lambda_0\} = G_x(\lambda_0).$$

For the weakened version, we have

$$M_x(u; \mathcal{S}) = \left[G_{x+1}^{-1}(u/2), G_x^{-1}((1+u)/2) \right),$$

so the new belief function is

$$\text{Bel}_x(\mathcal{A}; \mathcal{S}) = \max\{2G_x(\lambda_0) - 1, 0\}.$$

Again, this IM is somewhat conservative for the interval assertion $\mathcal{A}$.

2.5. *Using IMs for inference.* In this section we describe how belief or plausibility functions can be used for point estimation and hypothesis testing. This will further highlight the importance of Theorem 1.

The plausibility function $\text{Pl}_x(\cdot; \mathcal{S})$ is more convenient for use in the inference problem. Indeed, for any $\mathcal{A} \subset \Theta$, $\text{Pl}_x(\mathcal{A}; \mathcal{S})$ measures the amount of evidence in the observed data x that does not contradict the claim "$\theta \in \mathcal{A}$." So belief/plausibility functions are similar to Fisher's p-value in the sense that both tools attempt to assign post-data measures of (un)certainty to claims about the parameter of interest. One advantage of plausibility functions is that their interpretation is easier. Specifically, plausibility functions measure

one's uncertainty about the claim "$\theta \in \mathcal{A}$" given data, while p-values measure the probability of an observed event given the claim is true—reasoning about θ is *direct* with plausibility but somehow *indirect* with p-values.

Another advantage of belief/plausibility functions is that they can easily be used to design classical inference tools that satisfy the usual frequentist properties. For clarity, we reformulate the result of Theorem 1 in terms of the plausibility function.

COROLLARY 1. *Under the conditions of Theorem 1,*

$$\mathsf{P}_\theta\{\text{Pl}_X(\mathcal{A};\mathcal{S}) \leq \alpha\} \leq \alpha, \quad \forall\, \theta \in \mathcal{A}. \tag{2.9}$$

Hypothesis testing. Consider a "null hypothesis" $H_0 : \theta \in \mathcal{A}$, where $\mathcal{A}$ is a subset of Θ. Then an IM-based counterpart to a frequentist testing rule is of the following form:

$$\text{reject } H_0 \text{ if } \text{Pl}_x(\mathcal{A};\mathcal{S}) \leq t \text{ for a specified threshold } t \in (0,1). \tag{2.10}$$

According to Corollary 1, if the PRS $\mathcal{S}$ is credible, then the probability of a Type I error for such a rejection rule is

$$\mathsf{P}_\theta\{\text{Pl}_X(\mathcal{A};\mathcal{S}) \leq t\} \leq t.$$

So in order for the test (2.10) to control the probability of a Type I error at a specified $\alpha \in (0,1)$, one should reject H_0 if the plausibility is $\leq \alpha$.

Interval estimation. Consider a sequence of assertions $\mathcal{A}_t = \{t\}$ as t ranges over Θ. Now, for a counterpart to a frequentist confidence region, define the plausibility region

$$\Pi_x(\alpha) = \{t : \text{Pl}_x(\mathcal{A}_t;\mathcal{S}) > \alpha\}. \tag{2.11}$$

Now the coverage probability of the plausibility region (2.11) is

$$\begin{aligned} \mathsf{P}_\theta\{\Pi_X(\alpha) \ni \theta\} &= \mathsf{P}_\theta\{\text{Pl}_X(\mathcal{A}_\theta;\mathcal{S}) > \alpha\} \\ &= 1 - \mathsf{P}_\theta\{\text{Pl}_X(\mathcal{A}_\theta;\mathcal{S}) \leq \alpha\} \geq 1-\alpha, \end{aligned}$$

where the last inequality follows from Corollary 1. Therefore, this plausibility region has at least the nominal coverage probability.

3. Efficiency and conditioning. In Section 2 it was shown that credibility of the PRS was of fundamental importance. Martin, Zhang and Liu (2010) argue, however, that credibility cannot be the only consideration. They define a second criterion—*efficiency*—and formulate the choice of PRS as a constrained optimization problem. But this approach is for a fixed a-equation. Here we focus on a first step that modifies the a-equation to simplify the construction of a credible and efficient PRS. After this initial dimension reduction step is taken, the construction of an efficient IM is just as in Martin, Zhang and Liu (2010).

3.1. *Motivation: dimension reduction.* In the examples in Section 2.4, the dimension of the parameter is the same as that of the a-variable. But in general these dimensions will not be the same; in particular, the dimension of the a-variable will often be larger than that of the parameter. It is intuitively clear that efficient prediction of a-variables becomes more difficult as the dimension grows—this is basically the *curse of dimensionality*. In fact, constructing efficient PRSs for high-dimensional $U^\star$ can be quite challenging; see Martin, Zhang and Liu (2010). Therefore, we propose an initial dimension reduction step to make construction of PRSs simpler and the resulting a-inference more efficient.

EXAMPLE 5 (Normal model, cont.). Suppose $X_1, \ldots, X_n$ are iid observations from a $\mathsf{N}(\mu, 1)$ model with common unknown mean $\mu \in \mathbb{R}$. Then there are n copies of the a-equation $X_i = \mu + U_i$. Stacking these n a-equations in vector notation we have $X = \mu 1_n + U$, where 1_n is an n-vector of unity, and $U \in \mathbb{R}^n$ is distributed as $\mathsf{N}_n(0, I)$. Straightforward application of the reasoning in Example 1 suggests that inference on μ be carried out by predicting the unobserved auxiliary variable $U^\star$ in $\mathbb{R}^n$. But efficient prediction of $U^\star$ would be challenging if n is large, so reducing the dimension of $U^\star$—ideally to one dimension—would be a desirable first step.

In a likelihood-based inferential framework, an obvious approach to avoid the difficulty of the previous example would be to first reduce the data to a sufficient statistic, the sample mean in this case, and construct a new a-equation. In Section 3.2 we develop a framework that justifies this sort of intuition. We should mention, however, that while there are similarities between our conditioning approach and data reduction via sufficiency, the two are fundamentally different; see Sections 3.2, 3.3, and 3.5.

3.2. *Conditional IMs.* The main goal of this section is to effectively reduce the dimension of the a-variable $U^\star$ to be predicted by *conditioning*.

This is summarized in Theorem 2 below. But first we define formally what it means for two a-equation/a-measure pairs—*a-pairs*—to be equivalent for inference on θ. Recall the definition of the basic belief function (2.2).

DEFINITION 2. Consider two general a-pairs, say

$$p_1(X) = a_1(U_1, \theta), \quad U_1 \sim \nu_1 \quad \text{and} \quad p_2(X) = a_2(U_2, \theta), \quad U_2 \sim \nu_2.$$

These two are said to be equivalent for inference on θ if the corresponding basic belief functions Bel_x^1 and Bel_x^2 are identical.

As a simple example, it is easy to check that

$$X = \mu + U, \quad U \sim \mathsf{N}(0,1) \quad \text{and} \quad X = \mu + \Phi{-}1(U), \quad U \sim \mathsf{Unif}(0,1)$$

are equivalent in the sense of Definition 2 for inference on μ in the normal mean problem of Example 1. In this example, between the a-pairs listed above at least, there is no reason to prefer one over the other. In more complex problems, however, the dimension of the a-variable may vary over the candidate a-pairs. Our main goal is to show, via a conditioning argument, that we may choose the one with the lowest a-variable dimension without any loss of information.

THEOREM 2. *Suppose that the basic a-equation $x = a(u, \theta)$ in (2.1) can be expressed in the form*

$$p_1(x) = a_1(\varphi_1(u), \theta) \quad \text{and} \quad p_2(x) = a_2(\varphi_2(u)), \tag{3.1}$$

for suitable mappings p_1, p_2, a_1, a_2, φ_1, and φ_2. Write $v_1 = \varphi_1(u)$ and $v_2 = \varphi_2(u)$ for the new a-variables, and assume that the mapping $u \mapsto (v_1, v_2)$ is one-to-one and does not depend on θ. In addition, assume that the conditional focal element

$$\widetilde{M}_x(v_1) = \{\theta : p_1(x) = a_1(v_1, \theta)\}$$

is non-empty with $\nu\varphi_1^{-1}$-probability 1. Then the original a-pair is equivalent, in the sense of Definition 2, to the a-pair with conditional a-equation

$$p_1(x) = a_1(v_1, \theta) \tag{3.2}$$

and conditional a-measure $\tilde{\nu}_p$ being the conditional distribution of V_1, given that $a_2(V_2)$ equals the observed value $p_2(x) = p$.

PROOF. The basic a-equation for the original a-pair is

$$\text{Bel}_x(\mathcal{A}) = \frac{\nu\{u : M_x(u) \subseteq \mathcal{A},\ M_x(u) \neq \emptyset\}}{\nu\{u : M_x(u) \neq \emptyset\}}.$$

The assumption that the conditional focal element is non-empty implies that the belief function corresponding to the conditional a-pair is

$$\widetilde{\text{Bel}}_x(\mathcal{A}) = \tilde{\nu}_p\{v : \widetilde{M}_x(v) \subseteq \mathcal{A}\} = \frac{\nu\{u : \widetilde{M}_x(u) \subseteq \mathcal{A},\ a_2(\varphi_2(u)) = p\}}{\nu\{u : a_2(\varphi_2(u)) = p\}},$$

where p is the observed value $p_2(x)$. It is easy to check that the numerator of $\text{Bel}_x(\mathcal{A})$ is

$$\nu\{u : \widetilde{M}_x(\varphi_1(u)) \subseteq \mathcal{A},\ \widetilde{M}_x(\varphi_1(u)) \neq \emptyset,\ a_2(\varphi_2(u)) = p\},$$

and since $\widetilde{M}_x(\varphi_1(u)) \neq \emptyset$ with $\nu\varphi_1^{-1}$-probability 1, this reduces to

$$\nu\{u : \widetilde{M}_x(\varphi_1(u)) \subseteq \mathcal{A},\ a_2(\varphi_2(u)) = p\}.$$

Likewise, the denominator of $\text{Bel}_x(\mathcal{A})$ is simply $\nu\{u : a_2(\varphi_2(u)) = p\}$. Taking the ratio we get exactly $\widetilde{\text{Bel}}_x(\mathcal{A})$, which proves the claim. □

REMARK 1. The dimension reduction will be achieved by conditioning since, generally, $\varphi_1(u)$ will be of much lower dimension than u itself; in fact, $\varphi_1(u)$ will frequently be a "sufficient statistic" type of quantity, so its dimension will be the same as θ rather than x. See Section 3.5 below.

The significance of Theorem 2 is that we obtain a new *conditional a-equation* (3.2) which, together with a *conditional a-measure* $\tilde{\nu}_p$, for $p = p_2(x)$, can be used to construct a *conditional IM* along the lines in Section 2.3. That is, first specify a PRS $\mathcal{S} = \mathcal{S}(V)$ for predicting the unobserved a-variable $V^\star = \varphi_1(U^\star)$, and define the conditional focal elements

$$\widetilde{M}_x(v; \mathcal{S}) = \bigcup_{v' \in \mathcal{S}(v)} \{\theta : p_1(x) = a_1(v', \theta)\}. \tag{3.3}$$

Then the corresponding belief function is given by

$$\widetilde{\text{Bel}}_x(\mathcal{A}; \mathcal{S}) = \tilde{\nu}_p\{v : \widetilde{M}_x(v; \mathcal{S}) \subseteq \mathcal{A}\}, \quad \mathcal{A} \subseteq \Theta. \tag{3.4}$$

Then this belief function may be used, as in Section 2.5, to make inference on the unknown parameter θ. We revisit our examples in Section 3.3.

The very same credibility condition (Definition 1) can be considered and, in the case where $\tilde{\nu}_p$ is independent of x, the desirable frequency properties for $\widetilde{\text{Bel}}_x(\mathcal{A};\mathcal{S})$ in (3.4) follow immediately from Theorem 1. Surprisingly, $\tilde{\nu}_p$ is indeed independent of x in a number of important examples. But more generally, we would like to extend the notion of credibility and the result of Theorem 1 to the case where the conditional a-measure $\tilde{\nu}_p$ indeed depends on the observed data x. This is the topic of Section 3.4 below.

3.3. *Examples.* Here we revisit those examples presented in Section 2.4 to demonstrate the dimension reduction technique in Theorem 2.

EXAMPLE 6 (Normal model, cont.). Suppose $X_1, \ldots, X_n$ is an iid sample from a $\mathsf{N}(\mu, \sigma^2)$ population. If $\sigma^2 = 1$ is known, then the basic a-equation can be written as

$$\overline{X} = \mu + \overline{U}, \quad \text{and} \quad X_i - \overline{X} = U_i - \overline{U}, \quad i = 1, \ldots, n.$$

For an independent sample $U_1, \ldots, U_n$ from $\mathsf{N}(0,1)$, the sample mean $\overline{U}$ is independent of $U_i - \overline{U}$, $i = 1, \ldots, n$, and distributed as $\mathsf{N}(0, n^{-1})$. Therefore, there is no need to predict the entire n-vector of unobserved a-variables; IMs can be constructed for inference on μ based on predicting a single a-variable with *a priori* distribution $\mathsf{N}(0, n^{-1})$. More generally, for inference on (μ, σ^2), the basic a-equation reduces to

$$\{\overline{X} = \mu + \overline{U},\ s^2(X) = \sigma^2 s^2(U)\} \quad \text{and} \quad \frac{X_i - \overline{X}}{s(X)} = \frac{U_i - \overline{U}}{s(U)},$$

where $s^2(x) = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2$ denotes the sample variance. The vector of "z-scores" on the far-right of the previous display is known as the *sample configuration*, and is ancillary (maximal invariant) in general location-scale problems. Since $\overline{U}$ and $s^2(U)$ are independent, IMs for (μ, σ^2) can be built based on predicting $V^\star = (\overline{U}^\star, s^2(U^\star))$ using draws from the respective marginal distributions. The problem of, say, inference on μ alone when both μ and σ^2 are unknown is the topic of Part II of the series.

EXAMPLE 7 (Exponential model, cont.). For an iid sample $X_1, \ldots, X_n$ from an $\mathsf{Exp}(\lambda)$ population, the basic a-equation can be rewritten as

$$T(X) = \lambda T(U), \quad \text{and} \quad \frac{X_i}{T(X)} = \frac{U_i}{T(U)}, \quad i = 1, \ldots, n,$$

where $T(x) = \sum_{i=1}^n x_i$ is the sample total. The general result of Fraser (1966) shows that $T(U)$ and the vector $\{U_i/T(U) : i = 1, \ldots, n\}$ are independent,

so an IM for inference on λ can be built based on predicting $V^\star = T(U^\star)$ with draws from its marginal distribution $\mathsf{Gam}(n, 1)$.

EXAMPLE 8 (Bernoulli model, cont.). Suppose $X_1, \ldots, X_n$ is an iid sample from a $\mathsf{Ber}(\theta)$ population. An obvious extension of the model in Example 3 is to take $X_i = I_{\{U_i \le \theta\}}$, where $U_1, \ldots, U_n$ are iid $\mathsf{Unif}(0, 1)$. It turns out, however, that the conditioning approach does not easily apply to the model in this form. Here we consider an alternative formulation of the sampling model that leads to a very simple conditioning argument.

- Sample T, the total number of successes, by drawing $U_0 \sim \mathsf{Unif}(0, 1)$ and defining T such that

$$F_{n,\theta}(T) \le U_0 < F_{n,\theta}(T + 1), \tag{3.5}$$

  where $F_{n,\theta}(t)$ denotes the distribution function of $\mathsf{Bin}(n, \theta)$.
- Given T, randomly allocate the T successes and $n - T$ failures among the n trials. That is, randomly sample $(U_1, \ldots, U_n)$ from the subset of $\{0, 1\}^n$ consisting of exactly T ones, and set $X_i = U_i$, $i = 1, \ldots, n$.

We have defined the sampling model so that the decomposition (3.1) is explicit: the first step involves the parameter θ and a lower-dimensional summary of the data T and a-variable U_0, and the second is void of θ. Therefore, according to Theorem 2, we may consider the conditional a-equation (3.5), and it is easy to verify (via Bayes theorem) that the conditional a-measure—the distribution of U_0 given $(U_1, \ldots, U_n)$—is still $\mathsf{Unif}(0, 1)$. Therefore, a-inference on θ can proceed by predicting the unobserved $U_0^\star$ in the conditional a-equation (3.5). Note that this is exactly the result obtained by Zhang and Liu (2010) when only a single $T \sim \mathsf{Bin}(n, \theta)$ is observed.

EXAMPLE 9 (Poisson model, cont.). Suppose $X_1, \ldots, X_n$ are independent observations from a $\mathsf{Poi}(\lambda)$ population. We follow the approach described in Example 8 to construct a sampling model that leads to a simple conditioning argument.

- Sample T, the sample total, by drawing $U_0 \sim \mathsf{Unif}(0, 1)$ and defining T such that

$$F_{n\lambda}(T) \le U_0 < F_{n\lambda}(T + 1), \tag{3.6}$$

  where $F_{n\lambda}(t)$ is the distribution function of $\mathsf{Poi}(n\lambda)$.
- Randomly allocate portions of the total to the observations $X_1, \ldots, X_n$. That is, sample $(U_1, \ldots, U_n)$ from a $\mathsf{Mult}(T; n^{-1}1_n)$ distribution and set $X_i = U_i$, $i = 1, \ldots, n$.

Again, like in Example 8, the decomposition (3.1) of the sampling model into two components is made explicit. By Theorem 2 we may drop the second part, which does not involve λ, and consider the conditional a-equation (3.6). Similarly, it is easy to check that the conditional a-measure is simply $\mathsf{Unif}(0,1)$. Therefore, inference on λ can be carried out by predicting the unobserved $U_0^\star$ in the conditional a-equation (3.6).

The first two examples describe models which have a location and scale structure, respectively. (Example 8 has a similar, but less obvious, type of symmetry.) We show in Section 4.2 that Theorem 2 can be applied in problems that have a more general group transformation structure.

REMARK 2. The reader will have noticed that in each of the four examples presented above, the dimension reduction corresponds to that obtained by working with the data summarized by a sufficient statistic. However, the decomposition (3.1) is not unique, and the choice to use the familiar sufficient statistics is simply for convenience. For example, in Example 6, for inference on μ based on $X_1, \ldots, X_n$ iid $\mathsf{N}(\mu, 1)$, rather than using the conditional a-equation $\overline{X} = \mu + \overline{U}$, we could have equivalently used $X_1 = \mu + U_1$ but with conditional a-measure being the distribution of U_1 given $\{U_i - U_1 : i = 2, \ldots, n\}$. By choosing the average $\overline{U}$ we can make use of the well-known facts that $\overline{U}$ has a normal distribution and is independent of the residuals $\{U_i - \overline{U} : i = 1, \ldots, n\}$. The point is that data reduction via sufficiency singles out only an important special case of the decomposition (3.1) used to construct a conditional IM. See Section 3.5.

3.4. *Conditional credibility.* The goal of this section is to extend the notion of credibility and the results of Theorem 1 to the case where the conditional a-measure $\tilde{\nu}_p$ depends on x through $p = p_2(x)$.

Let $\mathcal{S}$ be a set-valued mapping and, following the development in Section 2.3, define the map

$$Q_p(v; \mathcal{S}) = \tilde{\nu}_p\{V : \mathcal{S}(V) \not\ni v\}.$$

This is just like the non-coverage probability in (2.6) but evaluated under the conditional distribution $\tilde{\nu}_p$ for V. Analogous to Definition 1 we define conditional credibility as follows.

DEFINITION 3. A PRS $\mathcal{S} = \mathcal{S}(V)$ for predicting $V^\star$ is *conditionally credible* at level α given $p_2(X) = p$, if

$$\tilde{\nu}_p\{V^\star : Q_p(V^\star; \mathcal{S}) \geq 1 - \alpha\} \leq \alpha. \tag{3.7}$$

For the focal elements $\widetilde{M}_x(v;\mathcal{S})$ in (3.3) and the corresponding belief function $\widetilde{\mathrm{Bel}}_x(\cdot;\mathcal{S})$, we have a *conditional version* of Theorem 1.

THEOREM 3. *Suppose $\mathcal{S}$ is conditionally credible for predicting $V^\star$ at level α, given $p_2(X) = p$, and that $M_x(V;\mathcal{S}) \neq \emptyset$ with $\tilde{\nu}_p$-probability 1 for all x such that $p_2(x) = p$. Then for any assertion $\mathcal{A} \subset \Theta$, $\widetilde{\mathrm{Bel}}_X(\mathcal{A};\mathcal{S})$ in (3.4), as a function of X, satisfies*

$$\mathsf{P}_\theta\{\widetilde{\mathrm{Bel}}_X(\mathcal{A};\mathcal{S}) \geq 1-\alpha \mid p_2(X) = p\} \leq \alpha, \quad \forall\, \theta \in \mathcal{A}^c. \tag{3.8}$$

PROOF. The proof follows that of Theorem 3.1 in Zhang and Liu (2010). Let θ denote the *true* value of the parameter, and assume $\theta \in \mathcal{A}^c$. Then

$$\begin{aligned}\widetilde{\mathrm{Bel}}_x(\mathcal{A};\mathcal{S}) &\leq \widetilde{\mathrm{Bel}}_x(\{\theta\}^c;\mathcal{S}) \\ &= \tilde{\nu}_p\{v : \widetilde{M}_x(v;\mathcal{S}) \not\ni \theta\} = Q_p(V^\star;\mathcal{S})\end{aligned}$$

for each x satisfying $p_2(x) = p$. Therefore, $\widetilde{\mathrm{Bel}}_x(\mathcal{A};\mathcal{S}) \geq 1-\alpha$ implies $Q_p(V^\star;\mathcal{S}) \geq 1-\alpha$ and, consequently, the conditional probability of the former can be no more than that of the latter. That is,

$$\mathsf{P}_\theta\{\widetilde{\mathrm{Bel}}_X(\mathcal{A};\mathcal{S}) \geq 1-\alpha \mid p_2(X) = p\} \leq \tilde{\nu}_p\{V^\star : Q_p(V^\star;\mathcal{S}) \geq 1-\alpha\},$$

and the result follows from the conditional credibility of $\mathcal{S}$. □

Corollary 2 below shows that under slightly stronger conditions the the long-run frequency property (3.8) holds *unconditionally*; the proof follows immediately from the law of total probability.

COROLLARY 2. *Suppose that the conditions of Theorem 3 hold for almost all p. Then (3.8) holds unconditionally, i.e.,*

$$\mathsf{P}_\theta\{\widetilde{\mathrm{Bel}}_X(\mathcal{A};\mathcal{S}) \geq 1-\alpha\} \leq \alpha, \quad \forall\, \theta \in \mathcal{A}^c.$$

3.5. *Relations to sufficiency and classical conditional inference.* Fisher's theory of sufficiency beautifully describes how the observed data can be reduced in such a way that no information about the parameter of interest θ is lost. In cases where the dimension reduction via sufficiency alone is unsatisfactory—i.e., the dimension of the sufficient statistic is greater than that of the parameter—there is an equally beautiful theory of conditional inference also due to Fisher but later built upon by others; see Reid (1995), Fraser (2004), and Ghosh, Reid and Fraser (2010) for reviews. We have already seen a number of instances where familiar things like sufficiency,

ancillarity, and conditional inference have appeared in the new approach. But while there are some similarities, there are also a number of important differences which we mention here.

(i) Fisherian sufficiency requires a likelihood function to define and justify the data reduction. There is no likelihood function in the proposed framework, so a direct reduction of the *observed data* cannot be justified. However, we have a-variables with valid distributions which we are allowed to manipulate more-or-less as we please, and for certain (convenient) manipulations of these a-variables, the familar data reduction via sufficiency emerges.

(ii) The proposed approach to dimension reduction is, in some sense, more general than that of sufficiency. The key feature is the conditional a-measure attached to the decomposition (3.1). In the proposed framework, almost any choice of decomposition is valid, and we see immediately the effect of our choice: a "bad" choice may have a complicated conditional a-measure, while a "good" choice might be much easier to work with. In other words, the conditional a-measures play the role of assigning a preference ordering to the various choices of decompositions (3.1).

(iii) In problems where the reduction via sufficiency is unsatisfactory (i.e., when the minimal sufficient statistic has dimension greater than that of the parameter), the classical approach is to find a suitable ancillary statistic to condition on. A number of well-known examples (see Ghosh, Reid and Fraser 2010) suggest that this task can be difficult in general. In the proposed system, the choice of decomposition (3.1) is not unique, but the procedure is basically automatic for each fixed choice.

(iv) Although they contain no information about the parameter of interest, ancillary statistics are important for conditional inference in the classical sense as they identify a relevant subset of the sample space to condition on. It turns out that there are actually two notions of "relevant subsets" in the proposed framework. The first is in the conditional a-measure: building an IM based on the conditional a-measure effectively restricts the distribution of the a-variable to a lower-dimensional subspace defined by the observed value of $a_2(\varphi_2(U))$. The second is in the conditional credibility theorem: the long-run frequency properties of the conditional IM are evaluated on the subset of the sample space defined by the observed value of $p_2(X)$.

4. Theory for group transformation models. In this section we derive some general properties of conditional IMs for a fairly broad class of models which are invariant under a suitable group of transformations. For the most part, our notation and terminology matches that of Eaton (1989). The analysis is different from, but shares some similarities with Fraser's

work on fiducial/structural inference in group invariant problems.

4.1. *Group transformation models.* Consider a special case where P_θ is invariant under a group $\mathscr{G}$ of transformations g mapping $\mathbb{X}$ onto itself. That is, there is a corresponding group $\overline{\mathscr{G}}$ of transformations mapping Θ onto itself such that if $X \sim \mathsf{P}_\theta$ and $g \in \mathscr{G}$, then $gX \sim \mathsf{P}_{\overline{g}\theta}$ for a suitable $\overline{g} \in \overline{\mathscr{G}}$. Here gx denotes the image of x under g. A number of popular models fit into this framework: location-scale data (e.g., normal, Student-t, gamma), directional data (e.g., Fisher-von Mises), and various multivariate/matrix-valued data (e.g., Wishart). Next we give some notation and definitions.

Suppose that $\overline{\mathscr{G}}$ is *transitive*; i.e., for any pair $\theta_1, \theta_2 \in \Theta$ there exists $\overline{g} \in \overline{\mathscr{G}}$ such that $\theta_1 = \overline{g}\theta_2$. Choose an arbitrary reference point $\theta_0 \in \Theta$. Then the model $X \sim \mathsf{P}_\theta$ may be written in a-equation form as

$$X = gU, \tag{4.1}$$

where $U \sim \nu \equiv \mathsf{P}_{\theta_0}$, and g is such that the corresponding $\overline{g}$ produces $\theta = \overline{g}\theta_0$.

For a point $x \in \mathbb{X}$, the orbit O_x of x with respect to $\mathscr{G}$ is defined as

$$O_x = \{gx : g \in \mathscr{G}\} \subseteq \mathbb{X};$$

that is, O_x is all possible images of x under $\mathscr{G}$. If $\mathscr{G}$ is transitive, then $O_x = \mathbb{X}$. A function $f : \mathbb{X} \to \mathbb{X}$ is said to be *invariant* if it is constant on orbits; i.e., if $f(gx) = f(x)$ for all x and all g. A function f is a maximal invariant if $f(x) = f(y)$ implies $y = gx$ for some $g \in \mathscr{G}$. A related concept is *equivariance*. A function $t : \mathbb{X} \to \Theta$ is equivariant if $t(gx) = \overline{g}\, t(x)$. Roughly speaking, an equivariant function preserves orbits in the sense that $gx \in \mathbb{X}$ is mapped by t to a point on the orbit of $t(x) \in \Theta$.

4.2. *Decomposition of the a-equation.* The examples in Section 3.3 illustrate that the method of conditioning can be useful for reducing the dimension of the a-variable $U^\star$ to be predicted. This reduction, however, requires existence of a decomposition (3.1) of the basic a-equation. It would, therefore, be useful to find problems where such a decomposition exists. Next we give one general result along these lines.

Let $t : \mathbb{X} \to \Theta$ be an equivariant function, and let $f : \mathbb{X} \to \mathbb{X}$ be a maximal invariant function, each with respect to the underlying groups $\mathscr{G}$ and $\overline{\mathscr{G}}$. Then a-equation (4.1) can be equivalently written as

$$t(X) = \overline{g}\, t(U) \quad \text{and} \quad f(X) = f(U), \tag{4.2}$$

and this implicitly defines the mappings p_j, a_j, and φ_j $(j = 1, 2)$ in Theorem 2. Then the conditional a-measure $\tilde{\nu}_p$ is the distribution of $t(U)$ when

$U \sim \nu = \mathsf{P}_{\theta_0}$, given the value $p = f(x)$ of the maximal invariant $f(U)$. We summarize this result as a theorem.

THEOREM 4. *If P_θ is invariant with respect to a group $\mathscr{G}$ of mappings on $\mathbb{X}$ and if the corresponding group $\overline{\mathscr{G}}$ on Θ is transitive, then there is an a-equation decomposition (4.2), defined by an equivariant function $t(\cdot)$ and a maximal invariant function $f(\cdot)$.*

If $t(\cdot)$ is a minimal sufficient statistic, and if the group operation on $t(\mathbb{X})$ induced by $\mathscr{G}$ is transitive—which is the case in Examples 6 and 7 where $\mathscr{G}$ is the translation and scaling group, respectively—then Fraser (1966) showed that $t(U)$ and $f(U)$ are independent, so $\tilde{\nu}_p = \nu = \mathsf{P}_{\theta_0}$. In such cases, there is no difficulty in constructing a conditional IM that satisfies the conditions of Theorem 1. When $t(U)$ and $f(U)$ are not independent, things are not so simple. Fortunately, there is a general expression for the exact distribution of $t(U)$ given $f(U)$ in group transformation problems, due to Barndorff-Nielsen (1983), which we discuss next.

4.3. *Conditional a-measures.* Consider a group invariant model where the parameter θ belongs to an open subset of $\mathbb{R}^k$. Assume also that the maximum likelihood estimate (MLE) of θ exists and is unique. In such cases, the MLE $\hat{\theta}(x)$ is an equivariant statistic (Eaton 1989, Theorem 3.2), so $t(\cdot)$ in (4.2) can be taken as $\hat{\theta}(\cdot)$. That is, the conditional a-equation is just

$$\hat{\theta}(x) = \overline{g}\hat{\theta}(u).$$

If $f(x)$ is a maximal invariant as in Section 4.2, then furthermore assume that the map $x \mapsto (\hat{\theta}(x), f(x))$ is one-to-one. Unlike the classical theory of invariant statistical models, however, in our context, the data x is fixed and the a-variable u varies. For this particular model, the a-measure is a fixed member of this invariant family, and the same maps that apply to x are applied to u as well. Therefore, using the notation of Section 3, our focus will be on the distribution of $V_1 = \hat{\theta}(U)$, the MLE of θ_0, a *known* quantity, given $V_2 = f(U)$. Under these conditions, Barndorff-Nielsen (1983) shows that the exact conditional distribution has a density

$$h(v_1 | v_2, \theta_0) \propto |j(v_1)|^{k/2} \exp\{\ell(\theta_0) - \ell(v_1)\}, \tag{4.3}$$

where $\ell(\theta) = \ell(\theta; v_1, v_2)$ and $j(\theta) = j(\theta; v_1, v_2)$ are, respectively, the log-likelihood function and the observed Fisher information matrix, evaluated at θ. Formula (4.3), called the "magic formula" in Efron (1998) and discussed in more detail in Reid (1995, 2003), is exact for some models, including group

transformation models, and is accurate up to at least $O(n^{-1})$ for many other models with v_1 a general (approximately) ancillary statistic.

So when $\hat{\theta}(U)$ and $f(U)$ are not independent, the conditional distribution depends on the observed value of $f(U)$ and, hence, on the data x. Consequently, the credibility results of Section 2 fail to hold (Fisher 1936). But in such cases, the new conditional credibility results presented in Section 3.4 are available to help one build a conditionally credible IM.

Notice that a posterior belief function can be constructed based on either the true conditional a-measure $\tilde{\nu}_p$ or its approximation based on (4.3). Since (4.3) is close to a normal density function, one would expect that the latter approximate posterior belief function might be easier to compute. Therefore, an interesting practical question is how fast does the difference between the two posterior belief functions vanish as $n \to \infty$? This question will be considered in more detail elsewhere.

5. Fisher's problem of the Nile. Suppose two independent samples, namely $X_1 = (X_{11}, \ldots, X_{1n})$ and $X_2 = (X_{21}, \ldots, X_{2n})$, are available from $\mathsf{Exp}(\theta)$ and $\mathsf{Exp}(1/\theta)$ populations, respectively. The goal is to make inference on the unknown $\theta > 0$. This is referred to as a "Problem of the Nile" based its connection to an applied problem of Fisher (1973) concerning the fertility of land in the Nile river valley. From a statistical point of view, this is an interesting example where the ML estimate is not sufficient, so sampling distributions conditioned on a suitable ancillary statistic are more appropriate for tests, confidence intervals, etc.

The construction of an IM for inference on θ proceeds as in Example 7 by first constructing conditional a-equations for each of the X_1 and X_2 samples. If the basic a-equations are

$$X_1 = \theta U_1 \quad \text{and} \quad X_2 = \theta^{-1} U_2,$$

where $U_1 = (U_{11}, \ldots, U_{1n})$ and $U_2 = (U_{21}, \ldots, U_{2n})$ are independent $\mathsf{Exp}(1)$ samples, then the conditional a-equations are

$$T(X_1) = \theta V_1 \quad \text{and} \quad T(X_2) = \theta^{-1} V_2, \tag{5.1}$$

where $V_j = T(U_j)$, $j = 1, 2$, and $T(x) = \sum_{i=1}^n x_i$. The conditional a-measures are just like in Example 7; that is, $V_1, V_2 \sim \mathsf{Gam}(n, 1)$.

But this naive reduction produces a two-dimensional a-variable (V_1, V_2) for inference on a scalar parameter θ. Efficiency can be gained by reducing the a-variable further, so we consider a slightly more sophisticated reduction step. In particular, consider

$$p_1(X) = \theta V_1 \quad \text{and} \quad p_2(X) = V_2, \tag{5.2}$$

where

$$p_1(X) = \sqrt{T(X_1)/T(X_2)} \qquad p_2(X) = \sqrt{T(X_1)T(X_2)}$$
$$V_1 = \sqrt{T(U_1)/T(U_2)} \qquad V_2 = \sqrt{T(U_1)T(U_2)}.$$

We now have an a-equation decomposition of the form (3.1) with two scalar a-variables—one connected to the parameter θ and the other fully observed. Therefore, by Theorem 2, an IM for θ may be built by predicting the unobserved a-variable $V_1^\star$ from its conditional distribution given the observed value of $V_2^\star$. It turns out that this conditional a-measure is a known distribution, namely a *generalized inverse Gaussian* distribution (Barndorff-Nielsen 1977). Its density function is of the form

$$f(v_1|v_2 = p) = \frac{1}{2v_1 K_0(2p)} \exp\{-p(v_1^{-1} + v_1)\}, \tag{5.3}$$

where $K_0(\cdot)$ is the modified Bessel function of the second kind. As a final simplifying step, write the conditional a-equation as

$$p_1(X) = \theta F_p^{-1}(U), \quad U \sim \mathsf{Unif}(0,1), \tag{5.4}$$

where F_p is the distribution function of V_1, given the observed value p of V_2, corresponding to the density in (5.3). Therefore, inference about θ can proceed by predicting the unobserved uniform variable $U^\star$ in (5.4).

The reader will no doubt recognize some of the aforementioned quantities from the usual likelihood-based approach to this problem. In particular, the following facts are mentioned by Ghosh, Reid and Fraser (2010):

- $p_1(X)$ is the maximum likelihood estimate of θ,
- $p_2(X)$ is an ancillary statistic, and
- the pair $(p_1, p_2)(X)$ is jointly minimal sufficient for θ.

But our approach is different from the usual likelihood-based approach in a number of ways. In particular, the conditioning argument described above is just a first step towards inference on θ; the next steps towards an IM are to choose a PRS for predicting the unobserved $U^\star$ in (5.4) and to calculate the corresponding posterior belief function.

6. Discussion. The theory of IMs gives a general framework in which posterior belief functions are produced for inference. Two important properties of the IM approach are that no prior distribution needs to be specified on the parameter space, and that the posterior belief functions are designed in such a way that desirable long-run frequency properties are realized. The

fundamental idea behind this approach is that inference about the parameter θ is equivalent to predicting the unobserved a-variable(s) $U^\star$. Zhang and Liu (2010) and Martin, Zhang and Liu (2010) give a general introduction to this IM approach, but their guidelines are fairly limited and there examples consider rather complicated PRSs for problems of moderate- to high-dimensions. In this paper we propose to simplify the approach of Zhang and Liu (2010) and Martin, Zhang and Liu (2010) by taking an intermediate conditioning step in which the dimension of the a-variable is reduced to make construction of a credible/efficient IM more manageable.

Throughout this development, a number of interesting open questions emerge. First, is it possible to say that one decomposition (3.1) is somehow "better" than another? The importance of this question from a practical point of view is clear, but an answer would also shed light on the connection between the proposed dimension reduction technique via conditioning and classical sufficiency. Second, is it possible to consider more general types of decompositions (3.1)? For example, can the component $p_2(X) = a_2(\varphi_2(U))$ be replaced by something more general, such as $c(X, \varphi_2(U)) = 0$? This form of decomposition is required for difficult problem of inference on the correlation in a bivariate normal model with known means and variances. Our current notion of conditional credibility cannot handle this level of generality, but an extension along these lines would make the conditioning approach more applicable, and may also help in understanding the various notions of "relevant subsets" in Section 3.5.

In the examples presented here, the conditional IM approach has a considerable overlap with the classical notion of dimension reduction via sufficiency, and we more-or-less recreate the classical solutions but in a different context. This was done intentionally and we argue that the similarities do not reflect poorly on the proposed approach. On the contrary, the fact that by making convenient choices of a-equations and PRSs we can recreate classical solutions suggests that an "optimal" IM-based solution can do no worse these solutions in a frequentist sense. Moreover, one should also keep in mind that the IM-based output has a posterior probability-like interpretation, i.e., the posterior plausibility function measures the amount of evidence in the observed data in favor of the assertion in question. Compare this to the interpretation of a Neyman-Pearson test of significance: both procedures can be designed to control the Type I error rates, but only the IM can simultaneously provide a post-data measure of uncertainty.

Efron (1998) states that Fisher's fiducial argument (or something like it) may be a big hit in the 21st century. There is definitely a difference between IMs and fiducial, but the two are similar in spirit. It remains to be seen if

this framework of IMs, beginning with Zhang and Liu (2010) and Martin, Zhang and Liu (2010) and developed further in this series of papers, has what it takes to fulfill Efron's prediction.

References.

Barndorff-Nielsen, O. (1977). Exponentially decreasing distributions for the logarithm of particle size. *Proc. R. Soc. Lond. A.* **353** 401–419.

Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70** 343–365. MR712023

Berger, J. O. (1984). The robust Bayesian viewpoint. In *Robustness of Bayesian analyses. Stud. Bayesian Econometrics* **4** 63–144. North-Holland, Amsterdam. With comments and with a reply by the author. MR785367

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Anal.* **1** 385–402 (electronic). MR2221271

Berger, J. O., Bernardo, J. M. and Sun, D. (2009). The formal definition of reference priors. *Ann. Statist.* **37** 905–938. MR2502655

Dawid, A. P. (1985). Calibration-based empirical probability. *Ann. Statist.* **13** 1251–1285. With discussion. MR811493

Dempster, A. P. (1966). New methods for reasoning towards posterior distributions based on sample data. *Ann. Math. Statist.* **37** 355–374. MR0187357

Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.* **38** 325–339. MR0207001

Dempster, A. P. (1968). A generalization of Bayesian inference. (With discussion). *J. Roy. Statist. Soc. Ser. B* **30** 205–247. MR0238428

Dempster, A. P. (2008). Dempster-Shafer calculus for statisticians. *Internat. J. of Approx. Reason.* **48** 265–277.

Eaton, M. L. (1989). *Group invariance applications in statistics.* Institute of Mathematical Statistics, Hayward, CA. MR1089423

Efron, B. (1998). R. A. Fisher in the 21st century. *Statist. Sci.* **13** 95–122. MR1647499

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23** 1–22. MR2431866

Efron, B. (2010). The future of indirect evidence. *Statist. Sci.* To appear.

Ermini Leaf, D., Hui, J. and Liu, C. (2009). Statistical inference with a single observation of $N(\theta, 1)$. *Pak. J. Statist.* **25** 571–586.

Ermini Leaf, D. and Liu, C. (2010). A weak belief approach to inference on constrained parameters: elastic beliefs. *Working paper.*

Fisher, R. A. (1925). Theory of statistical estimation. *Proc. Cambridge Philos. Soc.* **22** 200–225.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proc. Roy. Soc. A* **144** 285–307.

Fisher, R. A. (1935a). The fiducial argument in statistical inference. *Ann. Eugenics* **6** 391–398.

Fisher, R. A. (1935b). The logic of inductive inference. *J. Roy. Statist. Soc.* **98** 39–82.

Fisher, R. A. (1936). Uncertain inference. *Proc. Amer. Acad. Arts & Sci.* **71** 245–258.

Fisher, R. A. (1973). *Statistical methods and scientific inference*, 3rd ed. Hafner Press, New York. MR0346955

Fraser, D. A. S. (1961). The fiducial method and invariance. *Biometrika* **48** 261–280. MR0133910

Fraser, D. A. S. (1966). On sufficiency and conditional sufficiency. *Sankhyā Ser. A* **28** 145–150. MR0211519

Fraser, D. A. S. (2004). Ancillaries and conditional inference. *Statist. Sci.* **19** 333–369. With comments and a rejoinder by the author. MR2140544

Ghosh, J. K., Delampady, M. and Samanta, T. (2006). *An introduction to Bayesian analysis*. Springer, New York. MR2247439

Ghosh, M., Reid, N. and Fraser, D. A. S. (2010). Ancillary statistics: a review. *Statist. Sinica*. To appear.

Little, R. (2010). Calibrated Bayes, for statistics in general, and missing data in particular. *Statist. Sci.* To appear.

Martin, R., Zhang, J. and Liu, C. (2010). Dempster-Shafer theory and statistical inference with weak beliefs. *Statist. Sci.* **25** 72–87.

Reid, N. (1995). The roles of conditioning in inference. *Statist. Sci.* **10** 138–157. MR1368097

Reid, N. (2003). Asymptotics and the theory of inference. *Ann. Statist.* **31** 1695–1731. MR2036388

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR760681

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press, Princeton, N.J. MR0464340

Shafer, G. (1979). Allocations of probability. *Ann. Probab.* **7** 827–839. MR542132

Walley, P. (1987). Belief function representations of statistical evidence. *Ann. Statist.* **15** 1439–1465. MR913567

Walley, P. (1996). Inferences from multinomial data: learning about a bag of marbles. *J. Roy. Statist. Soc. Ser. B* **58** 3–57. With discussion and a reply by the author. MR1379233

Yager, R. and Liu, L., eds. (2008). *Classic works of the Dempster-Shafer theory of belief functions* **219**. Springer, Berlin. MR2458525

Zabell, S. L. (1992). R. A. Fisher and the fiducial argument. *Statist. Sci.* **7** 369–387. MR1181418

Zhang, J. and Liu, C. (2010). Dempster-Shafer inference with weak beliefs. *Statist. Sinica*. To appear.

Department of Mathematical Sciences
Indiana University-Purdue University Indianapolis
402 North Blackford Street, LD270
Indianapolis, IN 46202, USA
E-mail: rgmartin@math.iupui.edu

Institute of Statistical Science
Academia Sinica
Taipei, Taiwan
E-mail: jshwang@stat.sinica.edu.tw

Department of Statistics
Purdue University
250 North University Street
West Lafayette, IN 47907, USA
E-mail: chuanhai@stat.purdue.edu