GENERAL THEORY OF INFERENTIAL MODELS II. MARGINAL INFERENCE

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

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

This paper is a continuation of the authors' theoretical investigation of inferential model (IMs); see Martin, Hwang and Liu (2010). The fundamental idea is that prior-free posterior probability-like inference with desirable long-run frequency properties can be achieved through a system based on predicting unobserved auxiliary variables. In Part I, an intermediate conditioning step was proposed to reduce the dimension of the auxiliary variable to be predicted, making the construction of efficient IMs more manageable. Here we consider the problem of inference in the presence of nuisance parameters, and we show that such problems admit a further auxiliary variable reduction via *marginalization*. Unlike classical procedures that use optimization or integration, the proposed framework eliminates nuisance parameters via a set union operation. Sufficient conditions are given for when this marginalization operation can be performed without loss of information, and in such cases we prove that an appropriately constructed IM is calibrated, in a frequentist sense, for marginal inference. In problems where these sufficient conditions are not met, we propose a marginalization technique based on *parameter expansion* that leads to conservative marginal inference. The marginal IM approach is illustrated on a number of examples, including Stein's problem and the Behrens-Fisher problem.

1. Introduction. In statistical inference problems, it is often the case that only some components (or, more generally, some lower-dimensional functions) of the parameter vector θ are of interest. Linear regression, with $\theta = (\beta, \sigma^2)$, is one such example where primary interest is in the vector β of regression coefficients. Semiparametric problems (Bickel *et al.* 1998), such as the Cox proportional hazards model, form another important class of examples. More formally, suppose θ can be decomposed as $\theta = (\psi, \xi)$, where ψ is the parameter of interest and ξ is the *nuisance parameter*. The goal is to make inference on ψ in the presence of unknown ξ.

In these nuisance parameter problems, a modification of the classical likelihood framework is called for. Frequentists often opt for *profile* likelihood

AMS 2000 subject classifications: Primary 62A01; secondary 68T37, 62F99

Keywords and phrases: Bayes, Behrens-Fisher problem, belief functions, frequentist, marginal likelihood, Stein's paradox, predictive random sets

methods (e.g., Cox 2006, Ch. 7), where ξ is replaced by its conditional maximum likelihood (ML) estimate $\hat{\xi}(\psi)$. The effect is that the likelihood function involves only ψ, the parameter of interest, so point estimates and hypothesis tests can be constructed as usual. The downside, however, is that no uncertainty in ξ is accounted for when it is fixed at its ML estimate. A Bayesian-like alternative is the *marginal* likelihood approach, which assumes an *a priori* probability distribution for ξ. The marginal likelihood for ψ is obtained by integrating out ξ with respect to this distribution. Marginal likelihood inference effectively accounts for uncertainty in ξ, but difficulties arise from the requirement of a prior distribution for ξ. Indeed, a subjective prior may be difficult to elicit, suitable reference priors may not exist, and even if an acceptable prior is available, computation of the marginal likelihood can be challenging if ξ is high-dimensional and its prior distribution is not of a convenient form.

This is a difficult problem and neither the profile nor the marginal likelihood approach on its own, without any extra qualifications, is fully satisfactory. In some sense, a compromise between the frequentist and Bayesian approaches is needed. Progress along these general lines has been made recently through the concept of inferential models (IMs), starting with Zhang and Liu (2010) and later built upon by Martin, Zhang and Liu (2010). The fundamental idea in these two papers is that inference on an unknown parameter θ is equivalent to predicting an unobserved auxiliary variable drawn at random from a fully known *a priori* distribution. The practical consequences of this idea are two-fold:

- No prior is needed, and yet the inferential output (a belief function) is probabilistic in nature and conditioned on the observed data.
- Long-run frequency properties of IM-based decision procedures are shown to be completely determined by coverage probabilities of user-defined random sets for predicting this auxiliary variable.

That is, the IM framework automatically produces inferential output which is meaningful for the problem at hand and, at the same time, calibrated in a long-run frequency sense. Unfortunately, it can be difficult to find suitable random sets in moderate- to high-dimensional problems with the right coverage probabilities. Therefore, a natural idea is to reduce the dimension of the auxiliary variable as much as possible before attempting to predict it. In this series of papers we attempt to build, from the ground up, a general IM framework based on this idea of dimension reduction.

In Part I (Martin, Hwang and Liu 2010) we observe that, in many problems, some functions of the unobserved auxiliary variable are actually ob-

served, so it may not be necessary to predict the full auxiliary variable itself. For such problems, they propose a method of dimension reduction based on *conditioning*, and give general sufficient conditions under which this dimension reduction can be carried out without loss of information. They also prove a "conditional version" of the fundamental Theorem 3.1 of Zhang and Liu (2010), and draw parallels between this conditional IM framework and Fisher's notion of sufficiency and ancillarity (Fisher 1925, 1934, 1935) in group transformation models.

This conditioning step, reviewed in Section 3, will often reduce the dimension of the auxiliary variable to that of the parameter. But in marginal inference problems, where only parts of the full parameter are of interest, we can expect to reduce the dimension even further. Here, in Part II, we develop the IM framework for marginal inference problems based on a second dimension reduction technique. It turns out that if the model is "regular" (in the sense of Definition 1) then an easily interpretable (strong) marginal inference exists. In Section 4.1 we prove that the lower-dimensional relationship between data, parameter, and auxiliary variable obtained from the marginalization step is equivalent to the basic model for inference on ψ. This leads immediately to a system of marginal IMs, with interesting connections to marginal likelihood methods in classical statistics. Some simple normal distribution examples illustrate this approach.

When the sampling model is not regular, there are a number of ways to reason towards a (weak) marginal inference. In Section 4.2 we consider a general approach based on a concept of *parameter expansion* and illustrate the corresponding weak marginal inference in Stein's many-normal-means problem. The Behrens-Fisher problem is taken up in Section 5, where we show that a particularly convenient choice of IM recreates one of the most popular frequentist solutions due to Hsu (1938) and Scheffé (1970). Section 6 contains a brief discussion, and proofs of theoretical results are collected in Appendix A.

2. Sampling model and a-events. The sampling model P_θ, indexed by a parameter $\theta \in \Theta$, is a probability measure on the sample space $\mathbb{X}$ that encodes the joint distribution of the data vector $X = (X_1, \ldots, X_n)'$. The case where the individual X_i's take values in a more general space can be handled similarly. As in Martin, Zhang and Liu (2010) and Martin, Hwang and Liu (2010), we assume that P_θ can be constructed as follows. Take a more-or-less arbitrary auxiliary space $\mathbb{U}$, equipped with a fully known probability measure ν, and consider a pair of mappings

$$p : \mathbb{X} \times \Theta \to \mathbb{K} \quad \text{and} \quad a : \mathbb{U} \times \Theta \to \mathbb{K}. \tag{2.1}$$

The space $\mathbb{K}$ will be determined by the context—all that matters in theory is that both p and a map to the same space, so that (2.2) below makes sense. Occasionally we may write p_θ and a_θ for these maps when θ is fixed. Now fix $\theta \in \Theta$ and choose X to satisfy

$$p_\theta(X) = a_\theta(U), \quad \text{where} \quad U \sim \nu. \tag{2.2}$$

In other words, the sampling model for X given θ is determined by the a-measure ν and the a-equation (2.2). The a-equation (2.2) is more general than that in Zhang and Liu (2010) and Martin, Zhang and Liu (2010). They consider $p(x, \theta) \equiv x$ in (2.2), which is most easily seen as a data-generation mechanism; this is also the structural-equation version of Fraser (1968). The case $a(u, \theta) \equiv u$ reduces to the pivotal-equation version as in Dawid and Stone (1982). These two special cases cover many well-known models, and the general case was shown by Segal (1938) to cover all continuous sampling models; see also Barnard (1995). We will see the benefit of the more general a-equation when we consider marginal inference in Section 4.

For an alternative construction of the sampling model P_θ given the a-equation and a-measure, take any measurable $B \subset \mathbb{X}$ and define

$$\mathbb{U}_B(\theta) = \{u : p_\theta(x) = a_\theta(u) \text{ for some } x \in B\}, \quad \theta \in \Theta. \tag{2.3}$$

Then the sampling model satisfies

$$\mathsf{P}_\theta(B) = \nu\{\mathbb{U}_B(\theta)\}. \tag{2.4}$$

The set $\mathbb{U}_B(\theta)$ will be called an *a-event* and will help justify our reasoning towards inference in the nuisance parameter problem. There is also an interesting connection between a-events and likelihood.

REMARK 1. Fisher (1922) recognized the importance of the likelihood function for general statistical inference problems. But he emphasized that the likelihood function is not a probability distribution for θ. In fact, if all that was observed was "$X \in B$," then the left-hand side of (2.4), as a function of θ, is the likelihood function. Its interpretation is *postdictive* in the sense that it can be used to compare different explanations of the observed outcome "$X \in B$." The difficulty is that while $\mathsf{P}_\theta(B)$ is a probability, changing θ changes the underlying probability space, so the usual laws of probability do not hold. However, the probability space is fixed on the right-hand side of (2.4) as θ varies, so it may be possible to give likelihood a *predictive* interpretation along these lines. For us, this connection between likelihood and the a-events (2.3) will pay off when we consider the marginal inference problem in Section 4.

3. Quick review of a-inference. Belief functions and inferential models (IMs) for statistical inference on a parameter θ are presented in detail in Zhang and Liu (2010), Martin, Zhang and Liu (2010), and Martin, Hwang and Liu (2010). Here we give a quick review for completeness.

3.1. *Belief functions.* Belief functions (Dempster 1967; Shafer 1976) are similar to, but more general than, probability measures. The Dempster-Shafer (DS) theory for statistical inference constructs a belief function on Θ as follows. For observed data $X = x$, define the focal elements

$$M_x(u) = \{\theta : p(x, \theta) = a(u, \theta)\}, \quad u \in \mathbb{U}, \tag{3.1}$$

representing all those θ which are consistent with the observed data x and the particular a-variable u. One can view these focal elements as "inverted" a-events. Let $\mathcal{A}$ be a subset of Θ—called an *assertion* about the parameter θ. Then the basic belief function, evaluated at $\mathcal{A}$, is defined as

$$\text{Bel}_x(\mathcal{A}) = \frac{\nu\{u : M_x(u) \subseteq \mathcal{A},\ M_x(u) \neq \emptyset\}}{\nu\{u : M_x(u) \neq \emptyset\}}, \tag{3.2}$$

and can be roughly interpreted as the probability that the random set $M_x(U)$, for fixed x and $U \sim \nu$, falls completely inside $\mathcal{A}$. A related quantity is the plausibility function, defined as

$$\text{Pl}_x(\mathcal{A}) = 1 - \text{Bel}_x(\mathcal{A}^c), \tag{3.3}$$

and the *subadditivity* of Bel_x implies that $\text{Bel}_x(\mathcal{A}) \leq \text{Pl}_x(\mathcal{A})$ for all $\mathcal{A}$. Martin, Hwang and Liu (2010) show how tests of hypotheses, plausibility regions, etc can be built from the belief/plausibility functions.

3.2. *Weak beliefs and IMs.* The basic belief function (3.2) is not calibrated in a frequentist sense for all assertions $\mathcal{A}$; cf. Example 3.1 of Zhang and Liu (2010). A method of *weak beliefs* has been proposed by Zhang and Liu (2010) that shrinks the basic belief function just enough so that desirable long-run frequency properties are realized. Martin, Zhang and Liu (2010) call this shrunken belief function an *inferential model* (IM). Here is the basic idea. From (2.2), if x is the observed data and θ is the true value of the parameter, then surely

$$p(x, \theta) = a(U^\star, \theta),$$

where $U^\star$ is the unobserved auxiliary variable $U \sim \nu$. It is intuitively clear that if we know x and can accurately predict/guess $U^\star$, then inference on θ

is possible. Fiducial and DS approach the problem in this way, but try to predict $U^\star$ with a random draw $U \sim \nu$. Zhang and Liu (2010) and Martin, Zhang and Liu (2010) argue that this is overly optimistic. The method of weak beliefs, on the other hand, tries to predict $U^\star$ with a random set $\mathcal{S}(U)$ assumed to contain the random draw $U \sim \nu$. More formally, let $\mathcal{S} : \mathbb{U} \to 2^{\mathbb{U}}$ be a set-valued mapping with the property that $u \in \mathcal{S}(u)$ for all u. The belief that $U^\star$ belongs to $\mathcal{S}(U)$, clearly weaker than the "conventional" belief that $U^\star = U$, produces a focal element larger than that in (3.1) and, hence, a smaller belief function. Define

$$M_x(u; \mathcal{S}) = \bigcup_{u' \in \mathcal{S}(u)} M_x(u) \tag{3.4}$$

for $M_x(u)$ defined in (3.2). Clearly, $M_x(u; \mathcal{S})$ can be no smaller than $M_x(u)$. Likewise, the corresponding belief function—the IM—defined as

$$\text{Bel}_x(\mathcal{A}; \mathcal{S}) = \frac{\nu\{u : M_x(u; \mathcal{S}) \subseteq \mathcal{A},\ M_x(u; \mathcal{S}) \neq \emptyset\}}{\nu\{u : M_x(u; \mathcal{S}) \neq \emptyset\}} \tag{3.5}$$

can be no more than $\text{Bel}_x(\mathcal{A})$. In most cases, the denominator in (3.5) is equal to 1, but see Example 7 and Ermini Leaf and Liu (2010) for discussion of the more general case. An appropriate choice of the mapping $\mathcal{S}$ will produce *credible* IMs with desirable long-run frequency properties. See Theorems 1 and 3 in Martin, Hwang and Liu (2010) for more details. Next is a relatively simple example to illustrate the main ideas.

EXAMPLE 1 (Normal model). Suppose that X is a single observation from a normal population, $\mathsf{N}(\mu, \sigma^2)$, with known standard deviation σ but unknown mean μ. One choice of the basic a-equation is

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

where Φ is the cdf of $\mathsf{N}(0, 1)$. If x is the observed value of X, then the basic belief function Bel_x is the $\mathsf{N}(x, \sigma^2)$ probability measure, exactly the fiducial and (flat prior) Bayes answer. An IM may be constructed by predicting the unobserved $U^\star$ with a PRS of the form

$$\mathcal{S}(u) = [u - \gamma u, u + \gamma(1 - u)], \quad \text{for some } \gamma \in [0, 1].$$

Martin, Zhang and Liu (2010) give a formula for $\text{Bel}_x(\{\mu \leq \mu_0\}; \mathcal{S})$, and Zhang and Liu (2010) prove that this belief function has the desirable long-run frequency properties so long as $\gamma \in [1/2, 1]$.

3.3. *Conditional a-inference.* An important feature of Example 1 is that the dimension of the data equals that of the unknown parameter. But in many cases, particularly in the iid setting, the dimension n of the data is greater than that of the parameter. Accurately predicting a high-dimensional a-variable is difficult, and if inference on a comparatively low-dimensional parameter is the goal, then efficiency can be gained by first reducing the dimension of the a-variable. This is the topic of Part I. The key point here is that, in many cases, certain functions of $U^\star$ are observed so it may not be necessary to predict the entire a-variable.

For this dimension reduction, the primary assumption in Martin, Hwang and Liu (2010) is that the basic a-equation can be written as

$$p_1(x,\theta) = a_1(v_1,\theta) \quad \text{and} \quad p_2(x) = a_2(v_2), \tag{3.6}$$

where $v_i = \varphi_i(u)$, $i = 1, 2$. The second constraint carries no information about θ, but since the value of $V_2^\star = \varphi_2(U^\star)$ is effectively observed, it carries some information about $U^\star$, which we condition on. Moreover, $v_1 = \varphi_1(u)$ will be lower-dimensional, often the same dimension as the parameter. Martin, Hwang and Liu (2010) prove that for general group transformation models, this reduction can be made without any loss of information. Indeed, when the partition (3.6) exists, the basic a-equation/a-measure pair (2.2), and the conditional a-equation

$$p_1(X,\theta) = a_1(V_1,\theta),$$

with the corresponding a-measure being the conditional distribution of V_1, given $a_2(V_2) = p_2(x)$, are equivalent for inference on θ.

EXAMPLE 2 (Normal model, cont.). Suppose $X_1, \ldots, X_n$ are iid observations from a $\mathsf{N}(\mu, \sigma^2)$ population with unknown μ but known σ. Following the ideas in Example 1, we can write the basic a-equation in vector form as

$$X = \mu 1_n + \sigma U, \quad U \sim \mathsf{N}_n(0, I),$$

where 1_n is an n-vector of unity and I is the $n \times n$ identity matrix. From this a-equation it would appear that IM-based inference would require that we predict the entire unobserved n-vector of a-variables $U^\star$. However, as mentioned above, certain functions of $U^\star$ are observed, making it unnecessary to predict the full vector. Partition the a-equation as

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

Theorem 2 of Martin, Hwang and Liu (2010) shows that the above system of equations that are independent of μ can be effectively "ignored." In other

words, the original n-dimensional a-variable can be reduced to a scalar in the conditional a-equation

$$\overline{X} = \mu + \sigma\, n^{-1/2}\, \Phi^{-1}(U), \quad \text{where} \quad U \sim \mathsf{Unif}(0,1).$$

Clearly this problem is now identical to the one in Example 1.

REMARK 2. In what follows, we assume that this first conditioning step has been performed. That is, unless otherwise stated, when we refer to the basic a-equation and a-measure, we mean the one obtained after this initial dimension reduction step has been taken.

4. Marginal a-inference. The marginal inference problem boils down to one in which the assertions of interest contain statements about the parameter ψ only. This problem can be handled completely within the general framework described in Section 3. However, we can expect an overall gain in efficiency if we incorporate the marginal nature of the problem into the construction of the IM. Again, this efficiency gain is achieved by reducing the dimension of the a-variable to be predicted. The following important example motivates our investigation.

EXAMPLE 3 (Stein's problem). Suppose that $X_1, \ldots, X_n$ are independent observations with $X_i \sim \mathsf{N}(\mu_i, 1)$, $i = 1, \ldots, n$. The means $\mu_1, \ldots, \mu_n$ are unknown. In vector notation, the sampling model is

$$X \sim \mathsf{N}_n(\mu, I),$$

where X and μ are the n-vectors of observations and means, respectively, and I is the $n \times n$ identity matrix. Re-parametrize μ as (ψ, ξ), where $\psi = \|\mu\|$ is the length of μ and $\xi = \mu/\psi$ is the unit vector in the direction of μ. The goal is to make inference on ψ. For point estimation, note that the natural estimate $\|X\|^2$ of ψ^2 performs poorly on average. That is,

$$\mathsf{E}(\|X\|^2) = \psi^2 + n,$$

so $\|X\|^2$ can drastically overestimate ψ^2 when n is large. The fact that estimates which are component-wise optimal can perform poorly in the compound problem is commonly referred to as *Stein's paradox* (Stein 1956). Attempts to understand this phenomenon have led, at least in part, to the development of empirical Bayes methods and shrinkage estimation (Efron and Morris 1977; Robbins 1956, 1964).

The basic a-equation for this problem is

$$X = \psi\xi + U, \quad U \sim \mathsf{N}_n(0, I). \tag{4.1}$$

The parameter ξ is not of interest. In fact, there is no less information about ψ in (4.1) if we let ξ range freely over $\mathbb{S}_n$, the unit n-sphere. That is, write

$$X = \psi\xi + U, \quad \textit{for some } \xi \in \mathbb{S}_n. \tag{4.2}$$

In light of (4.2) we can "integrate out" ξ, leaving the following relationship between X, U and ψ only:

$$\|X - U\|^2 = \psi^2. \tag{4.3}$$

This reasoning will be made more formal in the following subsections. Fiducial, DS, and (flat-prior) Bayes take (4.3) as the baseline for inference on ψ. But, from an a-inference perspective, (4.3) suggests that inference on a scalar ψ still requires prediction of a n-vector U. Clearly, efficiency can be gained by reducing the dimension of the a-variable.

Here we present an apparently new approach in which a sort of marginal likelihood is obtained directly from the joint likelihood without any prior for ξ. The key ingredient is the connection between a-events and likelihood mentioned in Remark 1—integration of the likelihood is replaced by taking unions of a-events over ξ. For problems in which these unions are suitably regular, a *strong* marginal a-event exists; otherwise, a *weak* marginal a-event is available. We define what it means to be "regular," and discuss each of these two methods in turn.

4.1. *Strong marginal a-inference.* In this section we consider the most natural form of marginal inference within our framework. The basic idea is to set up a relationship similar to (2.2) between the data x, the a-variable u, and the parameter of interest ψ. This boils down to being able to partition the basic a-equation into two, one involving ψ and the other involving ξ. Models in which this partitioning is possible are called *regular*. For the a-equation $p(x, \theta) = a(u, \theta)$ in Section 3, write

$$p(x; \psi, \xi) = a(u; \psi, \xi) \tag{4.4}$$

to emphasize the fact that θ is made up of two distinct components, ψ and ξ. Remember that the a-equation (4.4) is assumed to have already been through the conditioning process of Section 3.3; see Remark 2.

Definition 1. A sampling model with a-equation (4.4) is called *regular* for inference on ψ if there exists mappings φ, $\bar{p}$, $\bar{a}$, and c such that

$$\bar{p}(x, \psi) = \bar{a}(\varphi(u), \psi) \quad \text{and} \quad c(u, x, \xi) = 0, \tag{4.5}$$

or, equivalently, the a-event $\mathbb{U}_x(\psi, \xi)$ in (2.3) can be written as

$$\mathbb{U}_x(\psi, \xi) = \{u : \bar{p}(x, \psi) = \bar{a}(\varphi(u), \psi)\} \cap \{u : c(u, x, \xi) = 0\}. \tag{4.6}$$

The examples will show that regularity is not an unnatural property. But there are models which are not regular; see Example 6 and Section 5.

Suppose the model in question is regular. Then the condition $c(u, x, \xi) = 0$ carries no information about the parameter of interest ψ. So the actual value of ξ is not important, only that there is at least one ξ that satisfies this constraint for the given x and u. That is, for inference on ψ, a-equation (4.5) ought to be equivalent, in some sense, to

$$\bar{p}(x, \psi) = \bar{a}(\varphi(u), \psi) \quad \text{and} \quad c(u, x, \xi) = 0 \quad \textit{for some } \xi. \tag{4.7}$$

In terms of a-events, we can define

$$\begin{aligned} \mathbb{U}_x(\psi) &= \bigcup_\xi \mathbb{U}_x(\psi, \xi) \\ &= \bigcup_\xi (\{u : \bar{p}(x, \psi) = \bar{a}(\varphi(u), \psi)\} \cap \{u : c(u, x, \xi) = 0\}) \\ &= \{u : \bar{p}(x, \psi) = \bar{a}(\varphi(u), \psi)\}, \end{aligned} \tag{4.8}$$

where the last equality requires some mild conditions on the c-constraint; see Theorem 1. We call $\mathbb{U}_x(\psi)$ the *marginal a-event*.

The last line in (4.8) looks similar to the definition of the basic a-event in (2.3) but with a different constraint. The new constraint, namely

$$\bar{p}(x, \psi) = \bar{a}(w, \psi), \tag{4.9}$$

is what we call the *marginal* a-equation; the corresponding *marginal* a-measure is $\bar{\nu} = \nu\varphi^{-1}$, the distribution of $W = \varphi(U)$ for $U \sim \nu$. Again note that the dimension of $w = \varphi(u)$ will generally be smaller than that of u, often the same as that of ψ. Consequently, credible/efficient prediction of $W^\star$ should be easier than that of $U^\star$.

Definition 2. Consider two sets of a-equations and a-measures, say

$$\begin{aligned} p_1(X; \psi, \xi) &= a_1(U_1; \psi, \xi), \quad U_1 \sim \nu_1 \quad \text{and} \\ p_2(X; \psi, \xi) &= a_2(U_2; \psi, \xi), \quad U_2 \sim \nu_2. \end{aligned}$$

These two are said to be *equivalent for marginal inference on* ψ if the corresponding basic belief functions Bel_x^1 and Bel_x^2 are identical for all assertions of the form $\mathcal{A} = \Psi_0 \times \Xi$, for $\Psi_0 \subset \Psi$.

Next is the main result of this section. For the proof, see Appendix A.1.

THEOREM 1. *If the sampling model is regular, and if for any x and u, there exists ξ such that $c(u, x, \xi) = 0$, then the basic a-equation and the marginal a-equation (4.9) are equivalent for marginal inference on ψ.*

REMARK 3. In Remark 1 we highlighted the connection between likelihood and probabilities of a-events. This same connection remains in the present context, only here we obtain a form of marginal likelihood. Note that this "marginal likelihood" is defined not by integrating the joint likelihood over ξ but, rather, by taking a union of the basic a-events over ξ.

Theorem 1 above is similar in spirit to Theorem 1 of Martin, Hwang and Liu (2010) in that it re-expresses the basic a-equation in terms of a lower-dimensional a-variable. However, the latter is a general result that does not depend on the inference problem at hand, while the former makes use of the fact that only a component of the full parameter is of interest. This is a special case of a more general idea—assertion-specific IMs—that one's approach to inference can be tailored to fit the problem of interest.

For marginal inference on ψ, start with the marginal a-equation (4.9) and construct basic marginal focal elements as described in Section 3.1; that is,

$$\overline{M}_x(w) = \{\psi : \bar{p}(x, \psi) = \bar{a}(w, \psi)\} \subset \Psi, \quad w \in \mathbb{W}, \tag{4.10}$$

where $\mathbb{W}$ is the image of $\mathbb{U}$ under φ. The basic marginal belief function, evaluated at $\mathcal{A} \subset \Psi$, is again just the probability, under $W \sim \bar{\nu}$, that $M_x(W)$ falls completely inside $\mathcal{A}$, i.e.,

$$\overline{\text{Bel}}_x(\mathcal{A}) = \frac{\bar{\nu}\{w : \overline{M}_x(w) \subseteq \mathcal{A}, \, \overline{M}_x(w) \neq \emptyset\}}{\bar{\nu}\{w : \overline{M}_x(w) \neq \emptyset\}}. \tag{4.11}$$

This basic marginal belief function can be suitably weakened by incorporating a PRS $\mathcal{S} = \mathcal{S}(W)$ exactly as in Section 3.2, thereby producing a marginal IM on Ψ, written as $\overline{\text{Bel}}_x(\cdot; \mathcal{S})$, for inference on ψ.

An important question is if, for suitable PRS $\mathcal{S}$, the marginal IM is calibrated in a frequentist sense. We summarize the affirmative answer in the following theorem. The proof closely follows that of Theorem 3.1 in Zhang and Liu (2010) but, for completeness, Appendix A.2 contains the relevant definitions and a sketch of the main ideas.

THEOREM 2. *Suppose, in addition to the conditions of Theorem 1, that $\mathcal{S} = \mathcal{S}(W)$ is α-credible for predicting the unobserved $W^\star = \varphi(U^\star)$, and that*

$\overline{M}_x(W;\mathcal{S}) \neq \emptyset$ with $\bar{\nu}$-probability 1 for all x. Then for any assertion $\mathcal{A} \subset \Psi$, the marginal belief function $\overline{\text{Bel}}_X(\mathcal{A};\mathcal{S})$, as a function of X, satisfies

$$\mathsf{P}_{(\psi,\xi)}\left\{\overline{\text{Bel}}_X(\mathcal{A};\mathcal{S}) \geq 1-\alpha\right\} \leq \alpha, \quad \forall\, (\psi,\xi) \in \mathcal{A}^c \times \Xi.$$

Next we give two relatively simple examples to illustrate the proposed approach to marginal inference.

EXAMPLE 4 (Normal model, cont.). Suppose $X_1, \ldots, X_n$ are iid observations from a $\mathsf{N}(\mu, \sigma^2)$ distribution, where both μ and σ are unknown. Starting from where we left off in Example 2, we have the a-equation

$$\overline{X} = \mu + S\, n^{-1/2}\, U_1/U_2 \quad \text{and} \quad S = \sigma U_2,$$

where the a-variables U_1 and U_2 are independent, with

$$U_1 \sim \mathsf{N}(0,1) \quad \text{and} \quad (n-1)U_2^2 \sim \mathsf{ChiSq}_{n-1}.$$

From this, constructing a marginal a-equation for μ or σ is fairly simple. For example, if μ is the parameter of interest, then (after a change of a-variable and a-measure) the marginal a-equation is

$$\overline{X} = \mu + S\, n^{-1/2}\, F_n^{-1}(U), \quad U \sim \mathsf{Unif}(0,1), \tag{4.12}$$

where F_n is the distribution function of the t_{n-1} distribution. Moving terms around in (4.12) reveals the usual t-statistic used for classical inference on μ when σ is unknown. But rather than using the sampling distribution of the t-statistic for inference on μ, we proceed by building an IM based on predicting the unobserved value $U^\star$ of the uniform variate U in (4.12). Since we have reduced the dimension of the a-variable to 1, constructing a credible and efficient IM is straightforward.

EXAMPLE 5 (Normal model, cont.). Again, suppose $X_1, \ldots, X_n$ are iid observations from a $\mathsf{N}(\mu, \sigma^2)$ distribution, where both μ and σ^2 are unknown. The goal is to make inference on the standardized mean $\rho = \mu/\sigma$. Consider an a-event representation of the a-equation in Example 4; that is,

$$\mathbb{U}_x(\mu,\sigma) = \left\{(u_1,u_2) : \overline{x} = \mu + \frac{s}{n^{1/2}}\,\frac{u_1}{u_2} \text{ and } s = \sigma u_2\right\}.$$

Taking a union over all (μ,σ) such that $\mu = \rho\sigma$ gives the marginal a-event

$$\mathbb{U}_x(\rho) = \bigcup_{(\mu,\sigma):\mu=\rho\sigma} \mathbb{U}_x(\mu,\sigma) = \left\{(u_1,u_2) : \frac{n^{1/2}\overline{x}}{s} = \frac{n^{1/2}\rho + u_1}{u_2}\right\}.$$

Extracting the marginal a-equation we get

$$\frac{n^{1/2}\overline{X}}{S} = \frac{n^{1/2}\rho + U_1}{U_2},$$

and we recognize that the random variable on the right-hand side has a non-central t-distribution, namely $\mathsf{t}_{n-1}(n^{1/2}\rho)$. Making one more change of a-variable and a-measure, we can rewrite this marginal a-equation as

$$n^{1/2}\overline{X}/S = F_{n,\rho}^{-1}(U), \quad \text{where} \quad U \sim \mathsf{Unif}(0,1).$$

A credible and efficient IM for inference on ρ can then be constructed based on a PRSs for an unobserved uniform a-variable $U^\star$.

EXAMPLE 6 (Stein's problem, cont.). The basic a-equation $X = \psi\,\xi + U$ in (4.1) can be written in the form

$$\|X - U\|^2 = \psi^2 \quad \text{and} \quad \frac{X-U}{\|X-U\|} = \xi.$$

This appears to be of an acceptable form, but it turns out that the nuisance parameter ξ is lurking within the former equation. By expanding the sum-of-squares on the left-hand side and using (4.1), we find that

$$X'X = \psi^2 + 2\psi\xi' U + U'U,$$

which clearly involves ξ. Therefore, this model is not regular, so there is no strong marginal a-inference in this case. We will revisit this example once more in Section 4.2.

4.2. *Weak marginal a-inference.* Strong marginal inference, in the sense of Section 4.1, may not be available for a given sampling model. That is, even though we can always write a basic marginal a-event

$$\mathbb{U}_x(\psi) = \{u : p(x;\psi,\xi) = a(u;\psi,\xi) \text{ for some } \xi\},$$

this may not reduce to the convenient form (4.8). But recall that the primary goal of marginalization is to reduce the dimension of the a-variable to be predicted, which can be accomplished under less restrictive conditions than in Section 4.1. The key concept here is that of *weakening* the a-event; that is, we can "approximate" the marginal a-event $\mathbb{U}_x(\psi)$ by another event $\widetilde{\mathbb{U}}_x(\psi)$ that has the convenient form (4.8). In other words, we strive for a nice trade-off between weakening and dimension reduction for efficient marginal inference on ψ.

DEFINITION 3. An a-event $\widetilde{\mathbb{U}}_x(\psi)$ is said to be *weaker* than the marginal a-event $\mathbb{U}_x(\psi)$ if $\widetilde{\mathbb{U}}_x(\psi) \supseteq \mathbb{U}_x(\psi)$.

According to Definition 3, any superset of $\mathbb{U}_x(\psi)$ is a weak a-event. The intuition is that being less precise about the a-variable U may allow more flexibility in dealing with the nuisance parameter ξ. Our approach to weakening is to systematically relax/ignore some of the constraints defining the basic marginal a-event $\mathbb{U}_x(\psi)$ via a method of *parameter expansion*. Proposition 1 makes this precise.

PROPOSITION 1. *Suppose that $\theta \mapsto (\psi, \xi)$ is one-to-one and that there exists a mapping $\omega = \omega(\xi)$ such that $\mathbb{U}_x(\psi)$ can be written as*

$$\mathbb{U}_x(\psi) = \bigcup_{\xi} \{u : \bar{p}(x; \psi, \omega(\xi)) = \bar{a}(\varphi(u); \psi, \omega(\xi)) \text{ and } c(u, x, \xi) = 0\}.$$

Then the a-event $\widetilde{\mathbb{U}}_x(\psi)$, given by

$$\widetilde{\mathbb{U}}_x(\psi) = \bigcup_{(\xi,\omega)\in\Xi\times\Omega} \{u : \bar{p}(x; \psi, \omega) = \bar{a}(\varphi(u), \psi, \omega) \text{ and } c(u, x, \xi) = 0\},$$

is weaker than $\mathbb{U}_x(\psi)$. Furthermore, if for every x and u there exists a ξ such that $c(u, x, \xi) = 0$, then $\widetilde{\mathbb{U}}_x(\psi)$ simplifies to

$$\widetilde{\mathbb{U}}_x(\psi) = \bigcup_{\omega} \{u : \bar{p}(x; \psi, \omega) = \bar{a}(\varphi(u); \psi, \omega)\}. \tag{4.13}$$

Parameter expansion, in this context, starts by reparametrizing the nuisance parameter ξ as a pair (ξ, ω) where, initially, ω is taken as a function of ξ. Then the basic marginal a-event $\mathbb{U}_x(\psi)$ is weakened by treating ω as a free parameter on its own, independent of ξ.

EXAMPLE 7 (Stein's problem, cont.). Following up on Example 6 we see that the basic marginal a-event can be written as

$$\mathbb{U}_x(\psi) = \bigcup_{\xi} \left\{ u : x'x = \psi^2 + 2\psi\xi'x + u'u \text{ and } \frac{x-u}{\|x-u\|} = \xi \right\}.$$

By replacing ξ with ω, an independent copy of ξ, in the first constraint, we construct a weak marginal a-event as in Proposition 1, (4.13):

$$\widetilde{\mathbb{U}}_x(\psi) = \bigcup_{\omega} \{u : x'x = \psi^2 + 2\psi\omega'u + u'u\}.$$

In terms of a-equations, we can conclude that

$$X'X = \psi^2 + 2\psi\omega'U + U'U, \quad U \sim \mathsf{N}_n(0, I), \quad \textit{for some } \omega.$$

But, as a function of $U \sim \mathsf{N}_n(0, I)$, the quantity $\psi^2 + 2\psi\omega'U + U'U$ has the same distribution for all ω on the unit n-sphere, namely the non-central chi-square distribution $\mathsf{ChiSq}_n(\psi^2)$. If F_{n,ψ^2} denotes its distribution function, then a simple change of a-variable and a-measure then shows that the weak marginal a-equation may be written as

$$X'X = F^{-1}_{n,\psi^2}(U), \quad U \sim \mathsf{Unif}(0, 1).$$

Therefore, a-inference can be done by predicting a single uniform random variable rather than a n-dimensional normal random vector. Note, however, that the marginal focal element

$$\overline{M}_x(u) = \{\psi : x'x = F_{n,\psi^2}(u)\}$$

is empty for u in a set of positive ν-probability, so conditions of the usual credibility theorem are not satisfied. The elastic belief approach of Ermini Leaf and Liu (2010) could be used here to construct an IM having the desired frequency properties.

Stein's problem is particularly well-suited for the parameter expansion approach to weakening, due to the fact that the distribution of $\psi^2 + 2\psi\omega'U + U'U$ is independent of ω. But this will not be the case in general; see Section 5. There, the union over ω will require more careful analysis.

Next we give some general remarks about the long-run frequency properties of weak marginal IMs. It should be intuitively clear that by weakening the constraints issued by the underlying sampling model, the focal elements would surely expand, thereby shrinking the belief function. The details are somewhat problem-specific, so here we present only some heuristics.

If $\omega \in \Omega$ were fixed and known, then marginal inference on ψ would would rely on a focal element of the form:

$$\overline{M}^{\omega}_x(u; \mathcal{S}) = \bigcup_{u' \in \mathcal{S}(u)} \{\psi : \bar{p}(x; \psi, \omega) = \bar{a}(\varphi(u'); \psi, \omega)\}.$$

Uncertainty about ω suggests a weakened focal element

$$\overline{M}^{\Omega}_x(u; \mathcal{S}) = \bigcup_{\omega \in \Omega} \overline{M}^{\omega}_x(u; \mathcal{S}). \tag{4.14}$$

Ideally, $\overline{M}_x^{\omega}(u;\mathcal{S})$ will be somehow monotone in ω for fixed x and u, so that the union $\overline{M}_x^{\Omega}(u;\mathcal{S})$ reduces to a fixed choice of ω. In Stein's problem (Example 7) the weakened focal elements are constant in ω. In the Behrens-Fisher problem presented in Section 5 we prove that the weakened focal elements correspond to one of two possible extreme ω-values, depending on group sample sizes. These weak marginal focal elements can be further weakened with PRSs to achieve the desirable long-run frequency properties. A loss of efficiency, however, is the price one pays for this initial "weakening by parameter expansion" step.

5. Behrens-Fisher problem. The Behrens-Fisher problem is one of the most fundamental problems in statistics (Scheffé 1970). The problem concerns inference on the difference between two normal means, based on two independent samples, when the standard deviations are completely unknown. It turns out that there are no exact tests/confidence intervals that do not depend on the order in which the data is processed. Various approximations are available, in particular those due to Scheffé (1970) and Welch (1938, 1947). For a review of these and other procedures, see Kim and Cohen (1998) and Dudewicz *et al.* (2007). Here we use the proposed marginal inference methodology to derive an IM for the Behrens-Fisher problem.

Suppose independent samples $X_{11},\ldots,X_{1n_1}$ and $X_{21},\ldots,X_{2n_2}$ are available from the populations $\mathsf{N}(\mu_1,\sigma_1^2)$ and $\mathsf{N}(\mu_2,\sigma_2^2)$, respectively. Summarize the data sets with $\overline{X}_k$ and S_k, $k=1,2$, the respective sample means and standard deviations. The parameter of interest is $\delta=\mu_2-\mu_1$. When σ_1 and σ_2 are known, or unknown but proportional, inference on δ is fairly straightforward. But in the general case there is no simple solution. The following subsections outline an approach based on marginal IMs.

5.1. *A first marginalization step.* The basic sampling model is of the location-scale variety, so the general results in Martin, Hwang and Liu (2010) suggest that we may immediately reduce to a lower-dimensional model based on the sufficient statistics. That is, we may begin our analysis with the basic a-equations

$$\overline{X}_k = \mu_k + \sigma_k\, n_k^{-1/2}\, U_{1k}, \quad \text{and} \quad S_k = \sigma_k U_{2k}, \quad k=1,2, \tag{5.1}$$

where the a-variables are independent and, for $k=1,2$,

$$U_{1k}\sim\mathsf{N}(0,1) \quad \text{and} \quad (n_k-1)U_{2k}^2\sim\mathsf{ChiSq}_{n_k-1}.$$

To incorporate $\delta=\mu_2-\mu_1$, combine the set of a-equations in (5.1) for μ_1 and μ_2 to get

$$\overline{D} = \delta + \sigma_2\, n_2^{-1/2}\, U_{12} - \sigma_1\, n_1^{-1/2}\, U_{11},$$

where $\overline{D} = \overline{X}_2 - \overline{X}_1$. Define $f(\sigma_1, \sigma_2) := [\sigma_1^2/n_1 + \sigma_2^2/n_2]^{1/2}$ and note that

$$\sigma_2\, n_2^{-1/2}\, U_{12} - \sigma_1\, n_1^{-1/2}\, U_{11}, \quad \text{and} \quad f(\sigma_1, \sigma_2)\, U_1, \quad U_1 \sim \mathsf{N}(0,1),$$

are equal in distribution. Therefore, making a change of a-variables leads to a new (and simpler) set of a-equations for the Behrens-Fisher problem:

$$\overline{D} = \delta + f(\sigma_1, \sigma_2)\, U_1, \quad \text{and} \quad S_k = \sigma_k U_{2k}, \quad k = 1, 2. \tag{5.2}$$

In terms of a-events, we may write

$$\begin{aligned}\mathbb{U}_x(\delta, \sigma_1, \sigma_2) &= \left\{(u_1, u_{21}, u_{22}) : \overline{d} = \delta + f(\sigma_1, \sigma_2)u_1,\ s_k = \sigma_k u_{2k}\right\} \\ &= \left\{(u_1, u_{21}, u_{22}) : \overline{d} = \delta + f(\tfrac{s_1}{u_{21}}, \tfrac{s_2}{u_{22}})u_1,\ s_k = \sigma_k u_{2k}\right\},\end{aligned}$$

and marginalizing over σ_1 and σ_2 gives

$$\mathbb{U}_x(\delta) = \{(u_1, u_{21}, u_{22}) : \overline{d} = \delta + f(\tfrac{s_1}{u_{21}}, \tfrac{s_2}{u_{22}})u_1\}. \tag{5.3}$$

We then pluck out the marginal a-equation for inference on δ:

$$\overline{D} = \delta + f(S_1/U_{21}, S_2/U_{22})U_1. \tag{5.4}$$

Since both data and a-variables are tied together in the f-function, in order to derive a strong marginal a-inference for δ, a factorization of the following form is needed:

$$f(s_1/u_{21}, s_2/u_{22}) = g_1(s_1, s_2) \times g_2(u_{21}, u_{22}). \tag{5.5}$$

Unfortunately, it does not appear that such a factorization exists, so we turn our attention to constructing a weak marginal a-inference for δ.

5.2. *Weak marginal a-inference.* Our approach is to "approximate" the factorization in (5.5). Towards this approximation, recall that the basic a-equation states that $s_k = \sigma_k u_{2k}$ for $k = 1, 2$. Under this constraint, the f-function satisfies

$$f^2(s_1/u_{21}, s_2/u_{22}) = f^2(s_1, s_2) \times \frac{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}{\frac{\sigma_1^2}{n_1}u_{21}^2 + \frac{\sigma_2^2}{n_2}u_{22}^2}.$$

The first factor involves data only; the second factor is data-free and depends only on σ_1, σ_2 and a-variables. In fact, the second factor is a function of a weighted average of the a-variables. Define the weight $\omega = \omega(\sigma_1, \sigma_2)$ as

$$\omega = \frac{\sigma_1^2/n_1}{\sigma_1^2/n_1 + \sigma_2^2/n_2} = \frac{1}{1 + n_1\sigma_1^2/n_2\sigma_2^2} \in \Omega := [0, 1].$$

If we consider ω as a function of (σ_1, σ_2), then the basic a-equation (5.2) can be rewritten as

$$\frac{\overline{D} - \delta}{f(S_1, S_2)} = \frac{U_1}{\sqrt{\omega U_{21}^2 + (1-\omega) U_{22}^2}} \quad \text{and} \quad \frac{n_1 S_2^2}{n_2 S_1^2} = \frac{U_{22}^2}{U_{21}^2} \frac{1-\omega}{\omega}. \tag{5.6}$$

The reader will recognize the quantity on the far left of (5.6) as the usual pivotal quantity used to construct confidence intervals, etc. The quantity ω, initially defined as a function of the nuisance parameter (σ_1, σ_2), will play the role of the expansion parameter as in Proposition 1.

Let $\mathbb{U}_x(\delta, \omega)$ denote the set of all (u_1, u_{21}, u_{22}) for which the constraints (5.6) hold; this is exactly $\mathbb{U}_x(\delta, \sigma_1, \sigma_2)$ above. Then the marginal a-event $\mathbb{U}_x(\delta)$ in (5.3) is just the union of $\mathbb{U}_x(\delta, \omega)$ over $\omega \in [0,1]$. Upon taking this union, the second constraint in (5.6) drops out, leaving just

$$\widetilde{\mathbb{U}}_x(\delta) = \bigcup_\omega \left\{ (u_1, u_{21}, u_{22}) : \frac{\overline{d} - \delta}{f(s_1, s_2)} = \frac{u_1}{\sqrt{\omega u_{21}^2 + (1-\omega) u_{22}^2}} \right\}. \tag{5.7}$$

It is clear that $\mathbb{U}_x(\delta) \subseteq \widetilde{\mathbb{U}}_x(\delta)$, so the latter is weaker than the former.

Equation (5.7) produces an "a-equation" of the form

$$\frac{\overline{D} - \delta}{f(S_1, S_2)} = \frac{U_1}{\sqrt{\omega U_{21}^2 + (1-\omega) U_{22}^2}} \quad \textit{for some } \omega.$$

But unlike in Stein's problem (Example 7), the distribution of the a-variable quantity on the right-hand side depends on ω. Let $G_\omega(z) = G_\omega(z; n_1, n_2)$ denote its distribution function. Then another change of a-variable gives

$$\frac{\overline{D} - \delta}{f(S_1, S_2)} = G_\omega^{-1}(U), \quad U \sim \mathsf{Unif}(0,1), \quad \textit{for some } \omega.$$

To construct an IM, take the simple PRS $\mathcal{S}(u) = [u/2, (1+u)/2]$ for predicting an unobserved uniform random variable. If ω were *known*, the IM under consideration would make use of the focal elements

$$\begin{aligned} \overline{M}_x^\omega(u; \mathcal{S}) &= \bigcup_{u' \in \mathcal{S}(u)} \left\{ \delta : \frac{\overline{d} - \delta}{f(s_1, s_2)} = G_\omega^{-1}(u') \right\} \\ &= \left\{ \delta : \frac{u}{2} \le G_\omega\Big(\frac{\overline{d} - \delta}{f(s_1, s_2)} \Big) \le \frac{1+u}{2} \right\}. \end{aligned}$$

Towards handling the additional uncertainty in ω, we present the following lemma, proved in Appendix A.3.

LEMMA 1. *Suppose X, Y_1, and Y_2 are independent random variables, with $X \sim \mathsf{N}(0,1)$ and $n_i Y_i \sim \mathsf{ChiSq}_{n_i}$, for $i = 1, 2$. For $\omega \in [0,1]$, define*

$$Z = \frac{X}{\sqrt{\omega Y_1 + (1-\omega) Y_2}},$$

and let $G_\omega(z)$ denote its distribution function. Then for all $\omega \in [0,1]$,

$$n_1 \leq n_2 \quad \textit{implies} \quad \begin{cases} G_\omega(z) \leq G_1(z), & z \leq 0 \\ G_\omega(z) \geq G_1(z) & z \geq 0. \end{cases}$$

Assume, without loss of generality, that $n_1 \leq n_2$. It follows from Lemma 1 and the symmetry of $G_\omega(z)$ about $z = 0$ for all ω that the weakened focal element (4.14) is

$$\overline{M}_x^\Omega(u; \mathcal{S}) = \bigcup_{\omega \in \Omega} \overline{M}_x^\omega(u; \mathcal{S}) = \overline{M}_x^1(u; \mathcal{S}).$$

That is, "conservative" marginal inference about ψ via parameter expansion corresponds to taking $\omega = 1$, i.e., the weak marginal a-equation is

$$\frac{\overline{D} - \delta}{f(S_1, S_2)} = G_1^{-1}(U), \quad U \sim \mathsf{Unif}(0,1).$$

Note that, in this context, G_1 is the distribution function of $\mathsf{t}_{n_1 \wedge n_2 - 1}$.

For the weak marginal IM described above, the plausibility can be easily calculated for any assertion about δ. For example, consider the assertion $\mathcal{A} = \{\delta = \delta_0\}$ for fixed $\delta_0 \in \mathbb{R}$. Then

$$\overline{\text{Pl}}_x(\mathcal{A}; \mathcal{S}) = 1 - \left| 2G_1\Big(\frac{\overline{d} - \delta_0}{f(s_1, s_2)}\Big) - 1 \right|, \tag{5.8}$$

and this can be used to perform tests of hypotheses or to construct plausibility intervals for δ as described in Martin, Hwang and Liu (2010).

5.3. *Example.* Data on travel times from home to work for two different routes are presented by Lehmann (1975, p. 83) and summarized in Table 1. The goal is to determine if the two routes have the same mean travel times. The plausibility function (5.8) for this data is shown in Figure 1. As expected, δ_0 values near the observed $\overline{d} = 1.444$ are highly plausible, and those far away are not. In particular, the plausibility for $\delta_0 = 0$ is 0.099; therefore, with a rule that "rejects" a proposed δ_0 value iff its plausibility is less than 0.05, we must "accept" $\mathcal{A} = \{\delta = 0\}$ in this case. But note

TABLE 1
Summary data for the example in Section 5.3.

Route	Sample size (n)	Sample mean ($\overline{x}$)	Sample variance (s^2)
1	5	7.580	2.237
2	11	6.136	0.073

that the inferential output here is more meaningful than that of an ordinary significance test. Indeed, the output $\overline{\text{Pl}}_x(\mathcal{A}; \mathcal{S}) = 0.099$ has a posterior probability-like interpretation measuring how likely it is, given the observed data, that δ equals 0. Compare this to the interpretation of a p-value.

A 95% plausibility interval for δ is obtained by finding all those δ_0 values such that the plausibility function (5.8) is greater than 0.05. After a bit of algebra, the plausibility interval is simply

$$\left\{\delta_0 : \overline{d} - f(s_1, s_2) G_1^{-1}(0.975) \leq \delta_0 \leq \overline{d} + f(s_1, s_2) G_1^{-1}(0.975)\right\}.$$

This region is displayed graphically in Figure 1; numerically, we have

$$95\% \text{ plausibility interval for } \delta: \quad (-3.314, 0.427).$$

It turns out that, for our particular choice of PRS $\mathcal{S}(u)$, the plausibility interval matches up exactly with the interval proposed by Hsu (1938) (see also Scheffé 1970) based on approximating the degrees of freedom in the t-distribution by $n_1 \wedge n_2 - 1$.

6. Discussion. In this paper we have considered the problem of inference in the presence of nuisance parameters. Classical approaches attempt to eliminate nuisance parameters via optimization or integration, but here the general idea is that unions of a-events over the nuisance parameters can be used to construct a-equations involving the parameter of interest and a lower-dimensional a-variable. This dimension reduction technique, in addition to that based on conditioning in Martin, Hwang and Liu (2010), simplifies the task of building a credible and efficient IM. For the case when the sampling model is regular (in the sense of Definition 1), marginal inference is straightforward, and according to Theorem 1 the dimension reduction is achieved without any loss of information or efficiency. In non-regular problems, however, a general approach to marginal inference has yet to emerge. We present one technique, based on an idea of parameter expansion, for constructing a weak marginal a-equation, and illustrations on Stein's example and the Behrens-Fisher problem are provided. Whether the proposed weak marginal IMs are efficient remains an open question; however, we conjecture

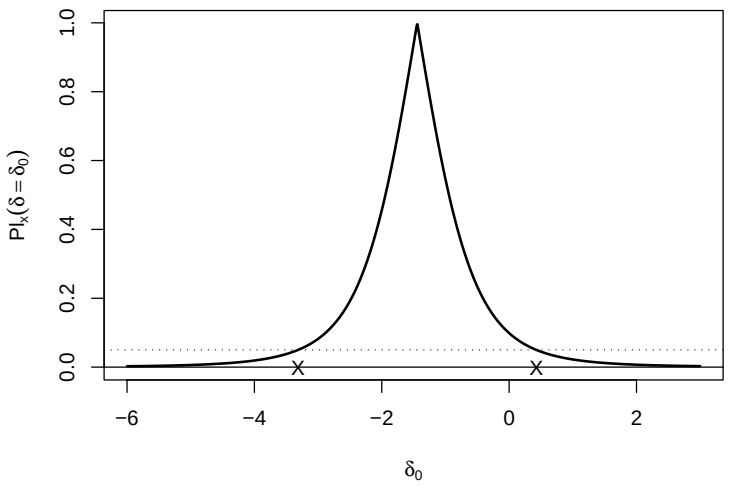

FIG 1. *Plot of the plausibility function* (5.8) *for the example in Section 5.3. The interval bounded by the two* X*'s is the 95% plausibility interval for* δ.

that efficiency holds in cases such as Stein's problem (Example 7) where the marginal a-measure does not depend on the expansion parameter.

In the context of marginal inference, the Behrens-Fisher problem is particularly challenging. The IM-based solution presented in Section 5 matches up exactly with one of the popular classical solutions (Hsu 1938; Scheffé 1970). But our derivation is based on a particularly convenient choice of PRS, and other choices would lead immediately to different solutions. Although the chosen PRS $\mathcal{S}_{1/2}(u)$ is "optimal" (Zhang and Liu 2010) in the class

$$\mathcal{S}_\gamma(u) = [u - \gamma u, u + \gamma(1-u)], \quad \gamma \in [0,1],$$

there is no reason to believe that this is the only reasonable class of PRSs. Therefore, the best IM-based solution to Behrens-Fisher problem can do no worse (in terms of Type I error rates or coverage probabilities) than existing solutions. But it is important to keep in mind that, in general, the conclusions of an IM-based analysis are more informative than that of, say, a Neyman-Pearson test of signifcance—plausibility values have a posterior probability-like interpretation in that they measure the amount of information in the observed data in favor of the assertion in question.

APPENDIX A: PROOFS

A.1. Proof of Theorem 1. In this context, the two sets of a-equations and a-measures in question are

$$p(X;\psi,\xi) = a(U;\psi,\xi), \quad U \sim \nu, \quad \text{and}$$
$$\bar{p}(X;\psi) = \bar{a}(W,\psi), \quad W \sim \nu\varphi^{-1}.$$

Fix $X = x$ and denote the corresponding basic belief functions as Bel_x and $\overline{\text{Bel}}_x$. Take any fixed $\Psi_0 \subset \Psi$. Then we have

$$\begin{aligned}
&\{(\psi,\xi) : p(x;\psi,\xi) = a(u;\psi,\xi)\} \subseteq \Psi_0 \times \Xi \\
&\qquad \iff \{(\psi,\xi) : \bar{p}(x,\psi) = \bar{a}(\varphi(u),\psi) \text{ and } c(u,x,\xi) = 0\} \subseteq \Psi_0 \times \Xi \\
&\qquad \iff \{\psi : \bar{p}(x,\psi) = \bar{a}(\varphi(u),\psi)\} \subseteq \Psi \text{ and } \{\xi : c(u,x,\xi) = 0\} \neq \emptyset \\
&\qquad \iff \{\psi : \bar{p}(x,\psi) = \bar{a}(\varphi(u),\psi)\} \subseteq \Psi_0
\end{aligned}$$

Since the first and last statements are equivalent, their respective probabilities (with respect to $U \sim \nu$) must be equal. These two probabilities are exactly $\text{Bel}_x(\Psi_0 \times \Xi)$ and $\overline{\text{Bel}}_x(\Psi_0)$, proving the claim. □

A.2. Proof of Theorem 2. Define the function

$$Q(w;\mathcal{S}) = \bar{\nu}\{w' : \mathcal{S}(w) \not\ni w\},$$

which is the probability that the PRS $\mathcal{S}(W)$ misses the target w when $W \sim \bar{\nu} := \nu\varphi^{-1}$. Then the PRS $\mathcal{S} = \mathcal{S}(W)$ is said to be α-credible if

$$\bar{\nu}\{w : Q(w;\mathcal{S}) \geq 1-\alpha\} \leq \alpha.$$

The key idea behind the proof is that if ψ is the true parameter value, then the marginal focal element $\overline{M}_x(w;\mathcal{S})$ misses ψ if and only if the PRS $\mathcal{S}(W)$ misses $W^\star$. So, if $\psi \notin \mathcal{A}$, then by monotonicity of the belief function

$$\begin{aligned}
\overline{\text{Bel}}_x(\mathcal{A};\mathcal{S}) &\leq \overline{\text{Bel}}_x(\{\psi\}^c;\mathcal{S}) \\
&= \bar{\nu}\{w : \overline{M}_x(w;\mathcal{S}) \not\ni \psi\} \\
&= \bar{\nu}\{w : \mathcal{S}(w) \not\ni W^\star\} = Q(W^\star;\mathcal{S}).
\end{aligned}$$

Note that the *a priori* distribution of $W^\star$ is $\bar{\nu}$; that is, allowing X to vary according to the sampling model is equivalent to varying $W^\star$ according to $\bar{\nu}$. Therefore, we get the desired result:

$$\mathsf{P}_{(\psi,\xi)}\Big\{\overline{\text{Bel}}_X(\mathcal{A};\mathcal{S}) \geq 1-\alpha\Big\} \leq \bar{\nu}\{W^\star : Q(W^\star;\mathcal{S}) \geq 1-\alpha\} \leq \alpha,$$

where the last inequality follows by the α-credibility assumption.

A.3. Proof of Lemma 1. First we show that the distribution function $G'_\omega(z)$ of Z^2 is minimized at $\omega = 1$ for all fixed $z \geq 0$ when $n_1 \leq n_2$. Then we show that this implies the claim. Start by noticing that

$$(T_1, T_2, T_3) := \frac{1}{n_1 Y_1 + n_2 Y_2 + X^2} (n_1 Y_1, n_2 Y_2, X^2)$$

is a $\mathsf{Dir}(\alpha_1 = n_1/2, \alpha_2 = n_2/2, \alpha_3 = 1/2)$ random vector. Therefore,

$$\begin{aligned} G'_\omega(z) &= \mathsf{P}\{Z^2 \leq z\} = \mathsf{P}\left\{\frac{X^2}{n_1 Y_1 + n_2 Y_2} \leq z\right\} \\ &= 1 - \mathsf{P}\left\{\tfrac{\omega}{n_1} T_1 + \tfrac{1-\omega}{n_2} T_2 < \tfrac{1}{z} T_0\right\} \\ &= 1 - \mathsf{P}\left\{(\tfrac{\omega}{n_1} + \tfrac{1}{z}) T_1 + (\tfrac{1-\omega}{n_2} + \tfrac{1}{z}) T_2 < \tfrac{1}{z}\right\} \\ &= 1 - K \int_0^{A(\omega)} \int_0^{B(\omega,t_1)} t_1^{\alpha_1 - 1} t_2^{\alpha_2 - 1} (1 - t_1 - t_2)^{-1/2} \, dt_2 \, dt_1, \end{aligned}$$

where

$$A(\omega) = \frac{1/z}{\omega/n_1 + 1/z}, \quad B(\omega, t) = \frac{A(\omega) - t}{A(\omega)\{1 + (1-\omega)z/n_2\}},$$

and $K = K(n_1, n_2)$ is a positive constant. Let $\partial = \frac{\partial}{\partial \omega}$ denote the differentiation operator. Applying Leibniz's rule for differentiating under the integral sign, and using the fact that $B(\omega, A(\omega)) = 0$, we obtain

$$\partial G'_\omega(z) = -K \int_0^{A(\omega)} t_1^{\alpha_1 - 1} B(\omega, t_1)^{\alpha_2 - 1} (1 - t_1 - B(\omega, t_1))^{-1/2} \partial B(\omega, t_1) \, dt_1.$$

Some tedious but straightforward algebra/calculus reveals that

$$\partial G'_\omega(z) = K z^{-1/2} A(\omega)^{\alpha_1} C(\omega)^{\alpha_2 - 1/2} H_\omega(z),$$

where

$$C(\omega) = \frac{1/z}{(1-\omega)/n_2 + 1/z},$$

$$H_\omega(z) = \int_0^1 s^{\alpha_1 - 1} (1-s)^{\alpha_2 - 1} \left[\frac{A(\omega) s}{n_1} - \frac{C(\omega)(1-s)}{n_2}\right] f_\omega(s) \, ds,$$

and

$$f_\omega(s) = \left[\frac{1-\omega}{n_1} + \left(\frac{\omega}{n_1} - \frac{1-\omega}{n_2}\right) A(\omega) s\right]^{-1/2}.$$

Let $\omega^\star = n_1/(n_1 + n_2)$. It is easy to check that $H_{\omega^\star}(z) = 0$; also,

$$H_0(z) \propto 1 - C(0) > 0 \quad \text{and} \quad H_1(z) \propto A(1) - 1 < 0.$$

We will now show that $H_\omega(z) \leq 0$ for $\omega^\star < \omega < 1$, which implies that $G'_\omega(z)$, for fixed $z \geq 0$, is decreasing in ω on the interval $[\omega^\star, 1]$. That $H_\omega(z) \geq 0$ for $0 < \omega < \omega^\star$ follows by symmetry.

For $\omega \in (\omega^\star, 1)$, the function $f_\omega(s)$ is analytic and strictly decreasing in s. Therefore, the coefficients $\delta_\omega(u)$, $u \in \mathbb{N}^+$, of its power series are strictly negative. Moreover, it follows from integration-by-parts that

$$I(u) := \int_0^1 s^{\alpha_1 - 1}(1-s)^{\alpha_2 - 1} \left[\frac{A(\omega)s}{n_1} - \frac{C(\omega)(1-s)}{n_2} \right] s^u \, ds > 0.$$

Therefore, $H_\omega(z) = \sum_{u=0}^\infty \delta_\omega(u) I(u) < 0$, as was to be shown.

We have shown that $G'_\omega(z)$ is minimized at $\omega = 1$ (and maximized at $\omega = 0$). That this also holds for $G_\omega(z)$, the actual distribution function in question, follows by symmetry. Indeed,

$$G'_\omega(z) = \mathsf{P}\{Z^2 \leq z\} = G_\omega(\sqrt{z}) - G_\omega(-\sqrt{z}) = 2G_\omega(\sqrt{z}) - 1,$$

so minimizing $G_\omega(z)$ for fixed $z \geq 0$ is equivalent to minimizing $G'_\omega(z)$ for a fixed but possibly different value of z. □

REFERENCES

Barnard, G. A. (1995). Pivotal models and the fiducial argument. *International Statistical Review* **63** 309–323.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1998). *Efficient and adaptive estimation for semiparametric models*. Springer-Verlag, New York. MR1623559

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge University Press, Cambridge. MR2278763

Dawid, A. P. and Stone, M. (1982). The functional-model basis of fiducial inference. *Ann. Statist.* **10** 1054–1074. With discussion. MR673643

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

Dudewicz, E. J., Ma, Y., Mai, E. and Su, H. (2007). Exact solutions to the Behrens-Fisher problem: asymptotically optimal and finite sample efficient choice among. *J. Statist. Plann. Inference* **137** 1584–1605. MR2339261

Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American* **236** 1419–127.

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

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A* **222** 309–368.

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. (1935). The logic of inductive inference. *J. Roy. Statist. Soc.* **98** 39–82.

Fraser, D. A. S. (1968). *The structure of inference*. John Wiley & Sons Inc., New York. MR0235643

Hsu, P. L. (1938). Contributions to the theory of "Student's" t-test as applied to the problem of two samples. In *Statistical Research Memoirs* 1–24. University College, London.

Kim, S.-H. and Cohen, A. S. (1998). On the Behrens-Fisher problem: A review. *Journal of Educational and Behavioral Statistics* **23** 356–377.

Lehmann, E. L. (1975). *Nonparametrics: statistical methods based on ranks*. Holden-Day Inc., San Francisco, Calif. MR0395032

Martin, R., Hwang, J.-S. and Liu, C. (2010). General theory of inferential models I. Conditional inference. *Working manuscript.*

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

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I* 157–163. University of California Press, Berkeley and Los Angeles. MR0084919

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.* **35** 1–20. MR0163407

Scheffé, H. (1970). Practical solutions of the Behrens-Fisher problem. *J. Amer. Statist. Assoc.* **65** 1501–1508. MR0273732

Segal, I. E. (1938). Fiducial distributions of several parameters with application to a normal system. *Proc. Camb. Phil. Soc.* **34**.

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

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I* 197–206. University of California Press, Berkeley and Los Angeles. MR0084922

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika* **29** 350–362.

Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika* **34** 28–35. MR0019277

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