Statistical Science
1999, Vol. 14, No. 1, 1–28

Integrated Likelihood Methods for Eliminating Nuisance Parameters

James O. Berger, Brunero Liseo and Robert L. Wolpert

Abstract. Elimination of nuisance parameters is a central problem in statistical inference and has been formally studied in virtually all approaches to inference. Perhaps the least studied approach is elimination of nuisance parameters through integration, in the sense that this is viewed as an almost incidental byproduct of Bayesian analysis and is hence not something which is deemed to require separate study. There is, however, considerable value in considering integrated likelihood on its own, especially versions arising from default or noninformative priors. In this paper, we review such common integrated likelihoods and discuss their strengths and weaknesses relative to other methods.

Key words and phrases: Marginal likelihood, nuisance parameters, profile likelihood, reference priors.

1. INTRODUCTION

1.1 Preliminaries and Notation

In elementary statistical problems, we try to make inferences about an unknown *state of nature* ω (assumed to lie within some set Ω of possible states of nature) upon observing the value $X = x$ of some random vector $X = \{X_1, \ldots, X_n\}$ whose probability distribution is determined completely by ω. If X has a density function $f(x|\omega)$, strong arguments reviewed and extended in Berger and Wolpert (1988) and Bjørnstad (1996) suggest that inference about ω ought to depend upon X only through the *likelihood function* $L(\omega) = f(x|\omega)$ which is to be viewed as a function of ω for the given data x.

James O. Berger is Arts and Sciences Professor and Robert L. Wolpert is Associate Professor at the Institute of Statistics and Decision Sciences, Duke University, Box 90251, Durham, North Carolina 27708-0251 (e-mail: berger@stat.duke.edu and rwolpert@stat.duke.edu). Brunero Liseo is Associate Professor at the Dip. di Studi Geoeconomici, Statistici e Storici per l'Analisi, Regionale, Università di Roma "La Sapienza" (e-mail: brunero@pow2.sta.uniroma1.it).

Rarely is the entire parameter ω of interest to the analyst. It is common to select a parameterization $\omega = (\theta, \lambda)$ of the statistical model in a way that simplifies the study of the "parameter of interest," here denoted θ, while collecting any remaining parameter specification into a "nuisance parameter" λ.

In this paper we review certain of the ways that have been used or proposed for eliminating the nuisance parameter λ from the analysis to achieve some sort of "likelihood" $L^*(\theta)$ for the parameter of interest. We will focus on integration methods, such as eliminating λ by simple integration (with respect to Lebesgue measure), resulting in the *uniform-integrated likelihood*

$$L^U(\theta) = \int L(\theta, \lambda)\, d\lambda. \tag{1}$$

In justifying integration methods, we will occasionally refer to alternative maximization methods, such as the *profile likelihood*

$$\hat{L}(\theta) = \sup_{\lambda} L(\theta, \lambda). \tag{2}$$

(Typically the sup over λ is achieved at some value $\hat{\lambda}_\theta$, which we will call the *conditional mle.*) However, no systematic discussion of nonintegration methods will be be attempted.

EXAMPLE 1. Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. normal random variables with mean μ and variance σ^2 ($N(\mu, \sigma^2)$). Suppose the parameter of interest is σ^2 while μ is a nuisance parameter. (Thus, in the above notation, $\theta = \sigma^2$ and $\lambda = \mu$.) Here, easy computations yield

$$\begin{aligned} L^U(\sigma^2) &= \int L(\sigma^2, \mu)\, d\mu \\ &= \int \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\} d\mu \\ &= \frac{1}{(2\pi\sigma^2)^{(n-1)/2}\sqrt{n}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}, \\ \hat{L}(\sigma^2) &= \sup_{\mu} L(\sigma^2, \mu) = L(\sigma^2, \bar{x}) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}; \end{aligned}$$

note that $\bar{x}$ is the conditional mle. Since proportionality constants do not matter for likelihoods, $L^U(\sigma^2)$ and $\hat{L}(\sigma^2)$ differ only in the powers of σ^2 in the denominators.

Notationally, we will write integrated likelihoods as

$$L(\theta) = \int L(\theta, \lambda)\pi(\lambda \mid \theta)\, d\lambda, \tag{3}$$

where $\pi(\lambda|\theta)$ is the "weight function" for λ. (In this paper, we consider only examples with continuous λ taking values in Euclidean space, and consider only integration with respect to Lebesgue measure.) It is most natural to use Bayesian language, and call $\pi(\lambda|\theta)$ the "conditional prior density of λ given θ," although much of the paper will focus on nonsubjective choices of $\pi(\lambda|\theta)$. Also, we will have occasion to refer to a prior density, $\pi(\theta)$, for the parameter of interest θ. As is commonly done, we will abuse notation by letting the arguments define the prior; thus $\pi(\theta)$ is the prior for θ, while $\pi(\lambda)$ would be the (marginal) prior for λ.

1.2 Background and Preview

The elimination of nuisance parameters is a central but difficult problem in statistical inference. It has formally been addressed only in this century since, in the nineteenth century Bayes–Laplace school of "Inverse Probability" (see, e.g., Zabell, 1989), the problem was not of particular concern; use of the uniform integrated likelihood $L^U(\theta)$ was considered "obvious."

With the Fisher and Neyman rejection of the Bayes–Laplace school, finding alternative ways to eliminate nuisance parameters was felt to be crucial. Student's derivation of the sampling distribution of the mean of a normal population when the variance is unknown and the derivation of the distribution of the sample correlation coefficient of a bivariate normal population (Fisher, 1915, 1921), are probably the first examples of a frequentist approach to the problem. Both were based on derivation of a pivotal quantity whose distribution is free of the nuisance parameters. Other famous examples include the Bartlett (1937) test of homogeneity of variances and the various solutions of the Behrens–Fisher problem (Fisher, 1935).

There have been numerous efforts to create a likelihood approach to elimination of nuisance parameters (Barnard, Jenkins and Winston, 1962). The beginnings of the "modern" likelihood school can perhaps be traced to Kalbfleisch and Sprott (1970, 1974), who proposed systematic study of a variety of methods for eliminating nuisance parameters (including integrated likelihood) and opened the way to a rich field of research.

Probably the simplest likelihood approach to eliminating nuisance parameters is to replace them with their conditional maximum likelihood estimates, leading to the profile likelihood in (2); this can then be used as an ordinary likelihood. Many examples of misleading behavior of the profile likelihood (Neyman and Scott, 1948; Cruddas, Cox and Reid, 1989) have given rise to various "corrections" of the profile, which aim to account for the "error" in simply replacing λ by a point estimate. Among the advances in this area are the modified profile likelihood (Barndorff-Nielsen, 1983, 1988) and the conditional profile likelihood (Cox and Reid, 1987). These methods were primarily developed to provide higher-order asymptotic approximations to (conditional) sampling distributions of statistics of interest, such as the maximum likelihood estimator or the likelihood ratio. As a by-product, these approximate distributions can be interpreted as likelihood functions for the parameter of interest and/or used in a frequentist spirit via tail area approximations. However, use of these methods tends to be restricted to rather special frameworks (e.g., exponential families or transformation groups). Excellent references include Reid (1995, 1996), Fraser and Reid (1989); see also Sweeting (1995a, b, 1996) for a Bayesian version of these approaches. A different way to adjust the profile likelihood, based on the properties of the score function, is developed in McCullagh and Tibshirani (1990).

Other likelihood approaches arise when one or more components of the sufficient statistics have marginal or conditional distributions which depend on θ, but not on λ. In these cases, such distributions

are often used as the likelihood for θ, and are called the "marginal likelihood" or the "conditional likelihood." Basu (1977) gives an interesting example of conflicting marginal and conditional likelihoods for the same problem, indicating that use of these techniques is likely to remain somewhat arbitrary.

Marginal and conditional likelihoods are special cases of the more general *partial likelihood* (Cox, 1975). If there exists a partition (y, z) of the data x, such that

$$(4) \quad f(x \mid \theta, \lambda) = h(x)\, f_1(y \mid \theta, \lambda) f_2(z \mid y, \theta)$$

or

$$(5) \quad f(x \mid \theta, \lambda) = h(x)\, f_1(y \mid \theta) f_2(z \mid y, \theta, \lambda),$$

then the terms hf_1 in (4) or hf_2 in (5) are ignored, and the remaining factor is taken as the partial likelihood for θ. Since the ignored term does depend on θ, there is some loss of information. Supporters of the approach suggest, however, that one loses only the information about θ which is inextricably tied with the unknown parameter λ. Basu (Ghosh, 1988, page 319) criticizes the idea of partial likelihood on the ground that it usually cannot be interpreted in terms of sampling distributions, and one is left only with the possibility of exploring the shape of the particular observed partial likelihood.

From a subjective Bayesian point of view, the problem has a trivial solution: simply integrate the joint posterior with respect to the nuisance parameters and work with the resulting marginal posterior distribution of θ. From a philosophical or foundational level, it would be difficult to add much to the fascinating articles (Basu, 1975, 1977) comparing subjective Bayesian and likelihood methods for elimination of nuisance parameters. There is, however, considerable resistance to general implementation of subjective Bayesian analysis, centering around the fact that elicitation of a subjective prior distribution for multiple parameters can be quite difficult; this is especially so for nuisance parameters, whose choice and interpretation are often ambiguous.

Even if one is not willing to entertain subjective Bayesian analysis, we feel that use of integrated likelihood is to be encouraged. The integration must then be with respect to default or noninformative priors. Our main goal will thus be to discuss and illustrate integrated likelihood methods based on different choices of conditional noninformative priors for the nuisance parameters.

In Section 2, we argue in favor of the use of integrated likelihood on the grounds of simplicity, generality, sensitivity and precision. In Section 3, we will illustrate the uses and interpretations of integrated likelihood in various conditional approaches to inference, ranging from the pure likelihood to the fully Bayesian viewpoints. Section 4 reviews the various integrated likelihoods that have been considered; these primarily arise from different definitions of conditional noninformative priors for the nuisance parameters. The last section focuses on criticisms and limitations of integration methods.

Throughout the paper, we will repeatedly use several examples to illustrate and compare different methods. Several of these examples (Examples 4 and 7, in particular) are simple to state but of near impossible difficulty to analyze, in the sense that default methods of any type are questionable. In these examples at least, we will thus be asking effectively, "What is the best method of doing the impossible?" While firm conclusions cannot then be forthcoming, we feel that consideration of such extreme examples can greatly aid intuition.

1.3 Some Philosophical Issues

In this section we discuss several issues that are somewhat tangential to the main theme, but relate to overall perspective.

1.3.1 *What is the likelihood function?* Bayarri, DeGroot and Kadane (1988) argued that there can be no unique definition of a "likelihood function," rejecting as incomplete the usual (and usually vague) definitions such as the one Savage (1976) attributes to Fisher: "probability or density of the observation as a function of the parameter."

Such definitions give no guidance as to how to treat the value of an unobserved variable (for example, a future observation z): should we condition on it, as we do for unknown parameters, leading to what Bayarri, DeGroot and Kadane (1988) call the "observed" likelihood $L_{\text{obs}}(\theta, z) \equiv f(x|\theta, z)$? Or should we treat it like other observations, leading to the "random variable" likelihood $L_{\text{rv}}(\theta, z) \equiv f(x, z|\theta)$? They argue that likelihood-based inference will depend on this choice and offer examples illustrating that either choice may be the best one in different examples.

Here, we avoid this difficulty by assuming that the problem begins with a specified likelihood function of the form

$$f(x, \theta^*, \lambda^* \mid \theta, \lambda),$$

where (as before) x is the observed value of the data vector X, and θ^* and λ^* are unobserved variables of interest and nuisance variables, respectively, having probability distributions as specified by f. The vector θ^* could include a future observation z or some "random effects" that are of interest, for example;

thus any variables with known (conditional) distribution are put to the left of the bar, while those without given distributions are placed on the right. Following Butler (1988) we recommend immediate removal (by integration) of the nuisance variable λ^*, passing to

$$f(x, \theta^* \mid \theta, \lambda) = \int f(x, \theta^*, \lambda^* \mid \theta, \lambda) d\lambda^*; \tag{6}$$

this then would be the likelihood we study, seeking to eliminate λ. [In most of our examples "θ^*" will be absent and so we will just write $f(x|\theta, \lambda)$.] This elimination of λ^* by integration is noncontroversial, and should be acceptable to most statistical schools. See Bjørnstad (1996) for additional discussion.

EXAMPLE 2. *Random effects.* Let X_i be independent $N(\mu_i, 1)$ random variables, with unobserved means μ_i drawn in turn independently from the $N(\xi, \tau^2)$ distribution; we wish to make inference about $\theta = (\xi, \tau^2)$, ignoring the nuisance parameter $\lambda^* = \mu = (\mu_1, \ldots, \mu_p)$ (the random effects). Here, (6) becomes (with no θ^* or λ present)

$$\begin{aligned} f(x \mid \theta) &= \int_{R^p} (2\pi)^{-p/2} \\ &\quad \cdot \exp\left(- \sum_{i=1}^{p} \frac{(x_i - \mu_i)^2}{2} \right) (2\pi\tau^2)^{-p/2} \\ &\quad \cdot \exp\left(- \sum_{i=1}^{p} \frac{(\mu_i - \xi)^2}{2\tau^2} \right) d\mu_1 \cdots d\mu_p \\ &= \left(2\pi(1+\tau^2)\right)^{-p/2} \exp\left(- \sum_{i=1}^{p} \frac{(x_i - \xi)^2}{2(1+\tau^2)} \right) \\ &\propto (1+\tau^2)^{-p/2} \\ &\quad \cdot \exp\left(- p\,[s^2 + (\bar{x} - \xi)^2]/2(1+\tau^2) \right), \end{aligned} \tag{7}$$

where $\overline{x}$ is the sample mean and $s^2 = \sum (x_i - \overline{x})^2 / p$.

Again, most statistical schools would accept elimination of μ by integration here; for instance, the usual alternative of maximizing over the unobserved parameters μ_i would lead instead to the profile likelihood

$$\begin{aligned} \hat{L}(\theta) &= \sup_{\mu \in R^p} (2\pi)^{-p/2} \\ &\quad \cdot \exp\left(- \sum_{i=1}^{p} \frac{(x_i - \mu_i)^2}{2} \right) (2\pi\tau^2)^{-p/2} \\ &\quad \cdot \exp\left(- \sum_{i=1}^{p} \frac{(\mu_i - \xi)^2}{2\tau^2} \right) \\ &= (4\pi^2\tau^2)^{-p/2} \exp\left(- \sum_{i=1}^{p} \frac{(x_i - \xi)^2}{2(1+\tau^2)} \right) \\ &\propto \tau^{-p} \exp\left(- p\,[s^2 + (\bar{x} - \xi)^2]/2(1+\tau^2) \right), \end{aligned} \tag{8}$$

which differs from $L(\theta)$ by having a singularity at $\tau = 0$ that strongly (and wrongly) suggests that *any* data support an inference that $\tau^2 \approx 0$. In situations such as this it is often suggested that one use a local maximum of the likelihood. Interestingly, no local maximum exists if $s^2 < 4$. Even if $s^2 \geq 4$, the local maximum is an inconsistent estimator of τ^2 as $p \to \infty$; for instance, if $\tau^2 = 3$, the local maximum will converge to 1 as $p \to \infty$. Figure 1 indicates the considerable difference between the likelihoods; graphed, for $p = 6$, $s^2 = 4$ and $\overline{x} = \xi$ are $L(\tau^2) = f(x|\tau^2,\ \xi = \overline{x})$ and $\hat{L}(\tau^2, \overline{x})$. Note that we have "maximized" over ξ to eliminate that parameter for display purposes. Had we integrated over ξ in $f(x|\theta)$, the difference would have been even more pronounced.

1.3.2 *The subjective Bayesian integrated likelihood.* Since integrated likelihood methods can be viewed in a Bayesian light, it is useful to review the purist Bayesian position. For subjective Bayesians there is no ambiguity concerning how to treat nuisance parameters: all inference is based on the joint probability distribution $f(x, \theta, \lambda)$ of all parameters and variables, whether or not observed, which can be constructed from any of the conditional data distributions along with appropriate subjective prior distributions. In problems without nuisance parameters, for example, the joint density is $f(x, \theta) = f(x|\theta)\pi(\theta)$, the product of the likelihood and the prior distribution $\pi(\theta)$ for θ; in this setting, Bayes' theorem is simply the conditional probability calculation that the posterior density for θ is

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int f(x \mid \theta)\pi(\theta) d\theta} \propto L(\theta)\pi(\theta). \tag{9}$$

In the presence of a nuisance parameter λ, a subjective Bayesian will base the analysis on a full prior $\pi^B(\theta, \lambda)$, which can also be factored as $\pi^B(\theta, \lambda) = \pi^B(\theta)\pi^B(\lambda|\theta)$, the product of the marginal and conditional prior densities of θ and λ (given θ), respectively. The subjective Bayesian still seeks $\pi(\theta|x)$ and would accept an integrated likelihood $L(\theta)$ if it satisfied $\pi(\theta|x) \propto L(\theta)\pi^B(\theta)$. It is easy to see that the only $L(\theta)$ which satisfies this relationship is given (up to a multiplicative constant) by

$$L^B(\theta) = \int f(x \mid \theta, \lambda)\pi^B(\lambda \mid \theta)\, d\lambda. \tag{10}$$

This thus defines the unique integrated likelihood that would be acceptable to a subjective Bayesian.

1.3.3 *Why eliminate nuisance parameters?* First, an explanation of the question: while, almost by definition, final inference about θ needs to be free

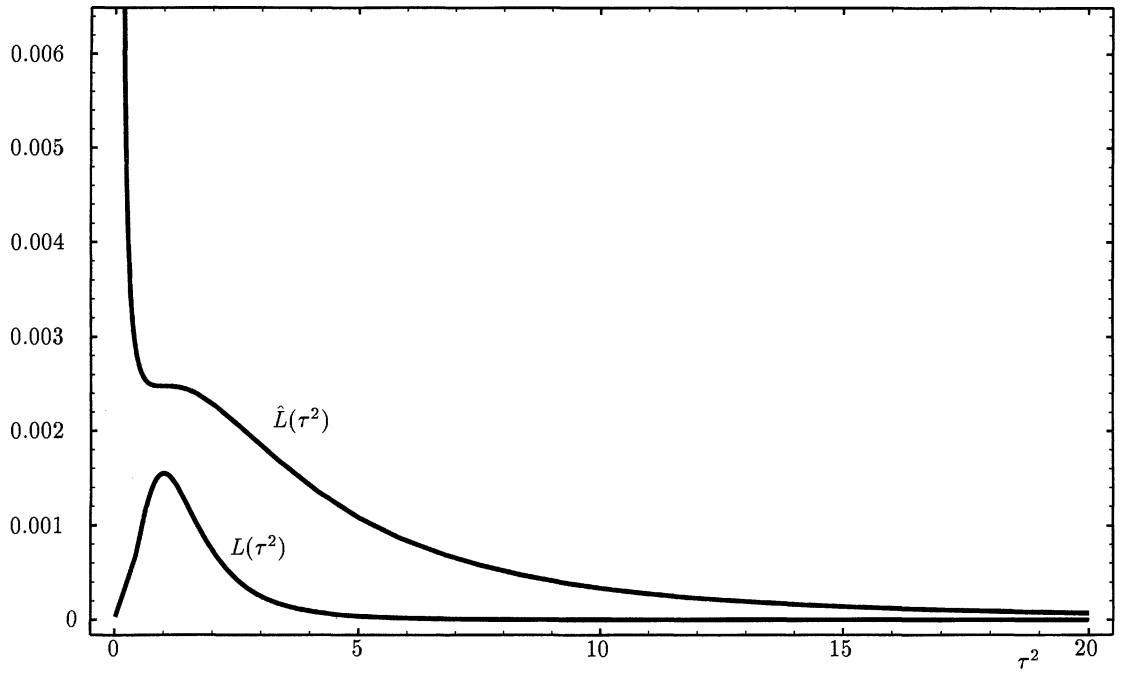

FIG. 1. *Integrated and profile likelihoods for the random effects model.*

of λ, it is not clear that one must pass through a likelihood function, $L(\theta)$, that is free of λ. Most non-Bayesian analyses pass through some such intermediary, but Bayesian analyses need not. For instance, many Bayesian analyses today proceed by Monte Carlo generation of a sequence of random variables $(\theta^{(1)}, \lambda^{(1)}), \ldots, (\theta^{(m)}, \lambda^{(m)})$ from the full posterior distribution $\pi(\theta, \lambda|x)$; inferences concerning θ then follow from direct use of the simulated values $\theta^{(1)}, \ldots, \theta^{(m)}$ (e.g., the usual estimate of θ would be the average of these simulated values). There is then no apparent need to explicitly consider an $L(\theta)$ that is free of λ. (Indeed, it is even common for Bayesians to introduce artificial nuisance parameters to simplify the Monte Carlo process.)

There are, nevertheless, several uses of $L(\theta)$ in Bayesian analysis, which we list here.

1. *Scientific reporting.* It is usually considered good form to report separately $L(\theta)$ and $\pi(\theta|x)$ (often graphically) in order to indicate the effect of the prior distribution. This also allows others to utilize their own prior distributions for θ.
2. *Sensitivity analysis.* It is often important to study sensitivity to $\pi(\theta)$, and having $L(\theta)$ available for this purpose is valuable. [Of course, sensitivity to $\pi(\lambda|\theta)$ is also a potential concern, but frequently this is of less importance.]
3. *Elicitation cost.* It is typically very expensive (in terms of time and effort) to obtain subjective prior distributions. Under the frequent cost limitations with which we operate, elicitation efforts may have to be limited. It is often cost effective to eliminate nuisance parameters in a default fashion, resulting in $L(\theta)$ and concentrate subjective elicitation efforts on $\pi(\theta)$.
4. *Objectivity.* Although most statisticians are justifiably skeptical of the possibility of truly "objective" analysis (cf. Berger and Berry, 1988), there is an undeniable need, in some applications, for an analysis which appears objective. Using $L(\theta)$, with default $\pi(\lambda|\theta)$, can satisfy this need.
5. *Combining likelihoods.* If one obtains information about θ from different independent sources, and the information arrives as likelihoods, $L_i(\theta)$, then one can summarize the information by $\prod_i L_i(\theta)$. This is the basis of many important meta-analysis techniques. One cannot, of course, simply multiply posteriors in this way. (But see Section 5.2 for cautions concerning multiplication of likelihoods.)
6. *Improper priors.* Focusing on integrated likelihood seems to reduce some of the dangers of using improper priors. This is illustrated in Section 3.2.

2. ADVANTAGES OF INTEGRATED LIKELIHOOD

Once one departs from the pure subjective Bayesian position, one cannot argue for integrated likelihood solely on grounds of rationality or co-

herency. Here we present a mix of pragmatic and foundational arguments in support of integrated likelihood. Other advantages will be discussed as we proceed.

2.1 Integration Versus Maximization

Most of the nonintegration methods are based on some type of maximization over the nuisance parameter. This can be very misleading if the likelihood has a sharp "ridge," in that the likelihood along this ridge (which would typically be that obtained by maximization) may be quite atypical of the likelihood elsewhere. Here is a simple example.

EXAMPLE 3. Suppose $X_1, \ldots, X_n$ are i.i.d. $N(\theta, 1)$ random variables, while Y is (independently) $N(\lambda, \exp\{-n\theta^2\})$. Here θ and λ are unknown, with θ the parameter of interest. The joint density for $X = (X_1, \ldots, X_n)$ and Y is

$$
\begin{aligned}
&f(x, y \mid \theta, \lambda) \\
&\quad = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right) \\
&\quad \cdot (2\pi \exp(-n\theta^2))^{-1/2} \exp\left(-\frac{(y-\lambda)^2}{2\exp(-n\theta^2)}\right) \qquad (11)\\
&\quad = (2\pi)^{-(n+1)/2} \\
&\quad \cdot \exp\left(-\frac{n}{2}(\overline{x}^2 - 2\overline{x}\theta) - \frac{(y-\lambda)^2}{2\exp(-n\theta^2)}\right),
\end{aligned}
$$

where $\overline{x}$ is the sample mean. For $n = 1$ and the data $x = 1$, $y = 0$, the overall likelihood $L(\theta, \lambda) = f(1, 0|\theta, \lambda)$ is graphed in Figure 2 for $\theta > 0$. Note the sharp ridge.

The profile likelihood is easy to compute, since the conditional mle for λ is just $\hat{\lambda}_\theta = y$. Thus

$$
\begin{aligned}
\hat{L}(\theta) &= f(x, y \mid \theta, \hat{\lambda}_\theta) \\
&\propto \exp(n\overline{x}\theta),
\end{aligned} \qquad (12)
$$

ignoring factors that do not involve θ. Note that this is a very strange "likelihood," rapidly growing to infinity as $\theta \to \infty$ or $\theta \to -\infty$, depending on the sign of $\overline{x}$.

In contrast, the uniform integrated likelihood is

$$
\begin{aligned}
L^U(\theta) &= \int f(x, y \mid \theta, \lambda)dx \\
&\propto \exp\left(-\frac{n}{2}(\overline{x} - \theta)^2\right),
\end{aligned} \qquad (13)
$$

which is clearly proportional to a $N(\overline{x}, 1/n)$ distribution, just as if Y with its entirely unknown mean λ had not been observed. These two likelihoods would, of course, give completely different impressions as to the location of θ.

The integrated likelihood answer can also be produced by a classical conditionalist. One can obtain the marginal likelihood for X by integrating out Y; the answer is clearly $L^U(\theta)$. There would thus be little disagreement as to the "correct" answer here, but the example does serve to indicate the danger inherent in maximization.

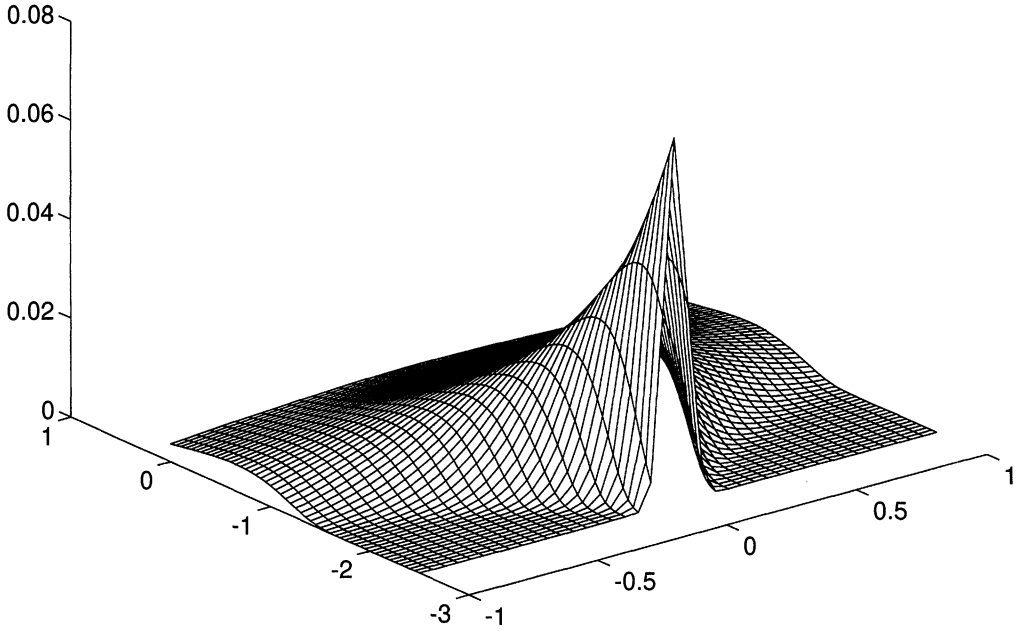

FIG. 2. *The likelihood surface for Example* 3, *when* $n = 1$, $x = 1$ *and* $y = 0$.

2.2 Accounting for Nuisance Parameter Uncertainty

The profile approach of replacing λ by its conditional mle, $\hat{\lambda}_\theta$, would appear to be dangerous in that it ignores the uncertainty in λ. Example 1 demonstrates a very mild version of this problem; the profile likelihood has one more "degree of freedom" than it should, given the replacement of μ by $\bar{x}$. We will mention a more serious standard example of this in Section 3.4, but do not dwell on the issue because it is a well-recognized problem. Indeed, many of the modifications to profile likelihood that have been advanced have, as one of their primary goals, adjustments to account for nuisance parameter uncertainty. It is important to emphasize that such modifications can be crucial, especially because "raw" likelihood procedures will typically be anticonservative, erring on the side of suggesting more accuracy than is actually warranted. Among the many disturbing examples of this is the exponential regression model (see discussion in Ye and Berger, 1991, and the references therein).

In contrast, integration methods automatically incorporate nuisance parameter uncertainty, in the sense that an integrated likelihood is an average over all the possible conditional likelihoods given the nuisance parameter. We do not claim that (default) integration methods are guaranteed to incorporate nuisance parameter uncertainty in a satisfactory way, but they certainly appear more naturally capable of doing so. As a final comment, note that if the conditional likelihoods do *not* vary significantly as the nuisance parameter λ changes, then the integrated likelihoods will be very insensitive to choice of $\pi(\lambda|\theta)$.

2.3 Simplicity and Generality

In comparing the simplicity of integration versus likelihood methods, it is difficult to draw firm conclusions because of the wide range of possible methods under each label. For instance, the profile likelihood in (2) is very simple to use while, on the integration side, the simplest is the uniform-integrated likelihood in (1). The various adjusted profile likelihoods and marginal likelihoods form an array of likelihood methods of modest to great complexity. Correspondingly, "optimal" integrated likelihoods can require difficult developments of noninformative priors. Computational considerations also come into play, although the historical wisdom that Bayesian computations are harder has today been reversed by the advent of MCMC computational techniques.

The key to comparison is judging the quality of the answers relative to the simplicity of the method. For instance, comparison at the simplest level, profile versus uniform-integrated likelihood, convinces many that the latter is considerably more effective in producing good answers in practice. Not many comparisons have been done at higher levels of complexity (Liseo, 1993, and Reid, 1996, are exceptions).

A few general observations are worth mentioning. First, integration methods are all based on the same basic idea; the only difference is in the prior distribution used to perform the integration. In contrast, the various profile, conditional and marginal likelihood approaches are based on very different rationales, and obtaining a feel for when each approach should be applied is not easy.

Second, the default Bayesian approaches have lately been used, with apparently great success, on a very large variety of complex applied problems. There are fewer successes in complex applied problems for the likelihood methods. This could, of course, be due to other factors, but is not irrelevant in terms of judging effectiveness versus difficulty.

A second relevant issue is generality of application. Likelihood methods are well known to have difficulties with nonregular problems, such as problems where the parameter is restricted to a certain range and the sample size is modest (so that likelihood surfaces can have strange shapes). Even worse is when the range restriction is affected by the data (e.g., a model where $x_i > \theta$, $i = 1, \ldots, n$), in which case standard asymptotics do not apply (Barndorff-Nielsen, 1991).

Another very difficult class of problems for likelihood methods is the Gleser–Hwang class, discussed in Section 5.1. A third difficult class is that consisting of problems involving discrete data and especially discrete parameters. These are "difficult" because the discreteness very much reduces the possibility of developing reasonable "adjustments" when basic methods are unreasonable. Here is a classic illustration.

EXAMPLE 4. *Binomial* (N, p). Consider k independent success counts $\mathbf{s} = (s_1, \ldots, s_k)$ from a binomial distribution with unknown parameters (N, p), and assume that N is the parameter of interest with p a nuisance parameter. This problem has been discussed in Draper and Guttman (1971), Carroll and Lombard (1985), Kahn (1987), Raftery (1988), Aitkin and Stasinopoulas (1989), Lavine and Wasserman (1992) and the references therein. Most of the points we make have already been made, in some form, in these articles.

The likelihood function is

$$L(N, p) = \left[\prod_{j=1}^{k} \binom{N}{s_j}\right] p^T (1-p)^{Nk-T}, \quad 0 < p < 1,\ N \geq s_{\max},$$

where $T = \sum_{j=1}^{k} s_j$ and $s_{\max} = \max_j s_j$. This likelihood is difficult to deal with by likelihood methods. For instance, the profile likelihood is

$$\hat{L}(N) = \left[\prod_{j=1}^{k} \binom{N}{s_j}\right] \frac{(Nk-T)^{Nk-T}}{(Nk)^{Nk}}, \quad N \geq s_{\max}, \tag{14}$$

and a "natural" conditional likelihood, the conditional distribution of $(s_1, \ldots, s_k)$ given T and N, is

$$L^C(N) = \left[\prod_{j=1}^{k} \binom{N}{s_j}\right] \Big/ \binom{kN}{T}, \quad N \geq s_{\max}. \tag{15}$$

For the data set $\mathbf{s} = (16, 18, 22, 25, 27)$, Figure 3 gives the graphs of $\hat{L}(N)$ and $L^C(N)$. These are nearly constant over a huge range of N and are clearly nearly useless for inference. Such behavior of $\hat{L}(N)$ and $L^C(N)$ is typical for this problem. [Indeed, $L^C(N)$ is often an increasing function of N.]

The uniform integrated likelihood is

$$L^U(N) = \int_0^1 L(N, p)\, dp = \left[\prod_{j=1}^{k} \binom{N}{s_j}\right] \frac{\Gamma(kN-T+1)}{\Gamma(kN+2)}, \quad N \geq s_{\max}. \tag{16}$$

This is also graphed in Figure 3 and appears to be much more useful than either $\hat{L}(N)$ or $L^C(N)$. There is more to be said here, and we will return to this example several times.

2.4 Sensitivity Analysis

One of the considerable strengths of using integrated likelihood is that one has a readily available sensitivity analysis: simply vary $\pi(\lambda|\theta)$, and see how $L^*(\theta)$ varies. This can be crucial in evaluating the robustness of the answer. Also, if considerable sensitivity is discovered, it is often possible to determine which features of $\pi(\lambda|\theta)$ are especially crucial, enabling either subjective elicitation of these features or the identification of additional data that could be collected to reduce sensitivity.

EXAMPLE 4 (Continued). Use of $L^U(\theta)$ corresponds to choice of a $U(0, 1)$ prior distribution for p. Another common noninformative prior is the Jeffreys prior, $\pi(p) \propto p^{-1/2}(1-p)^{-1/2}$, which will yield an integrated likelihood we denote by $L^J(\theta)$.

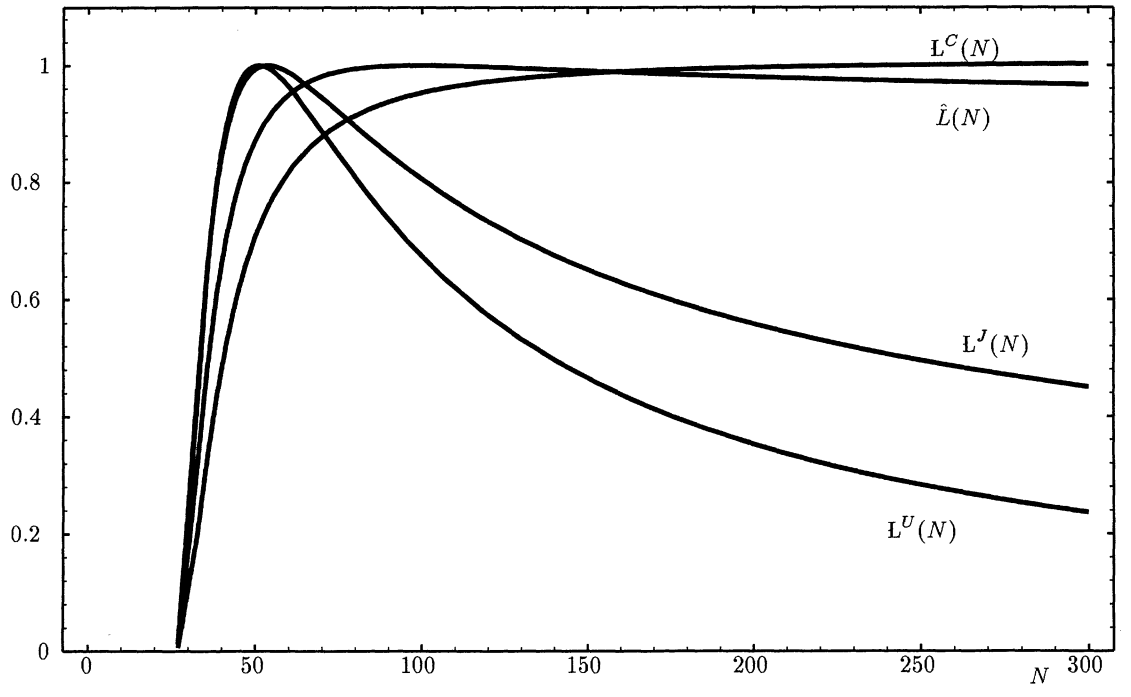

FIG. 3. *Likelihoods for N: $\hat{L}$ is the profile; L^C is the conditional; L^U is the uniform-integrated and L^J is the Jeffreys-integrated.*

More generally, one could consider Beta (a, a) prior densities for p; note that $a = 1$ and $a = 1/2$ yield the uniform and Jeffreys priors, respectively. Calculation yields, for the integrated likelihood with a Beta (a, a) prior for p,

$$L^a(N) = c_a\left[\prod_{j=1}^{k}\binom{N}{s_j}\right]\frac{\Gamma(kN-T+a)}{\Gamma(kN+2a)}, \quad N \geq s_{\max}, \tag{17}$$

where c_a is the prior normalization constant $c_a = \Gamma(2a)/(\Gamma(a))^2$.

We also graph $L^J(N) = L^{1/2}(N)$ in Figure 3; $L^a(N)$ for $1/2 < a < 1$ can be shown to lie between $L^J(N)$ and $L^U(N) = L^1(N)$, so that sensitivity over this range can be judged effectively simply by comparing L^J and L^U. A useful general result here is that

$$\begin{aligned}\frac{L^U(N)}{L^J(N)} &\approx \left(1-\frac{T+(0.7)}{kN+1}\right)^{1/2}\frac{\pi}{\sqrt{kN+1}} \\ &\approx \left(1-\frac{21.7}{N}\right)^{1/2}\frac{(1.4)}{\sqrt{N}} \quad \text{(in the example)},\end{aligned} \tag{18}$$

so that the main difference is the more quickly decreasing tail of L^U. It is worth emphasizing that all these reasonable "default" integrated likelihoods have tails that are considerably sharper than that of $\hat{L}(N)$ [or $L^C(N)$], indicating that the extremely flat tail of $\hat{L}(N)$ may be due to a "ridge" effect.

We later discuss the question of how to use integrated likelihoods. For now, we simply report the modes of $L^U(\theta)$ and $L^J(\theta)$ for the data $\mathbf{s}_1 = (16, 18, 22, 25, 27)$, along with those of $\hat{L}(\theta)$ and $L^C(\theta)$. Table 1 gives these modes along with the corresponding modes for two other data sets, $\mathbf{s}_2 = (16, 18, 22, 25, 28)$ and $\mathbf{s}_3 = (16, 18, 22, 25, 26)$. (The reason for considering such perturbations is that small errors in collection of count data such as these are almost inevitable, and one hopes the answer is not overly sensitive to such errors.)

TABLE 1
Modes of likelihoods for N

Likelihood type	Data set $\mathbf{s}_1$	$\mathbf{s}_2$	$\mathbf{s}_3$
Profile ($\hat{L}(N)$)	99	191	69
Conditional ($L^C(N)$)	∞	∞	∞
Uniform-integrated ($L^U(N)$)	51	57	46
Jeffreys-integrated ($L^J(N)$)	54	62	49

While modes alone are surely insufficient as a summary of likelihoods, the stability of those for the integrated likelihoods, over both change in the prior and small perturbations in the data, is quite appealing. For considerably more complete discussion of sensitivity of integrated likelihood in this problem, see Lavine and Wasserman (1992). We also should mention that sensitivity analysis of other types is certainly possible; see Olkin, Petkau and Zidek (1981) for an illustration.

3. INTERPRETATION AND USE OF INTEGRATED LIKELIHOOD

Since we are proposing use of integrated likelihood in general, and not only within the Bayesian paradigm, we need to discuss how it is to be interpreted and used. We discuss, in order, its use in likelihood analysis, Bayesian analysis and empirical Bayes analysis. Note that, of course, one might simply report the entire integrated likelihood function, leaving its interpretation and use to the consumer.

3.1 Use in Likelihood Analysis

Those likelihood methods which operate solely on the given $L(\theta)$ can also be used with an integrated likelihood. Examples of such methods are (1) using the mode, $\hat{\theta}$, of $L(\theta)$ as the estimate of θ; and (2) using (if θ is a p-dimensional vector)

$$C = \{\theta\colon\ -2\log(L(\theta)/L(\hat{\theta})) \leq \chi_p^2(1-\alpha)\}$$

as an approximate $100(1-\alpha)\%$ confidence set for θ, where $\chi_p^2(1-\alpha)$ is the $(1-\alpha)$th quantile of the chi-squared distribution with p degrees of freedom. The arguments which justify such methods for profile or modified profile likelihoods will typically also apply to integrated likelihoods (and can apply in more generality; see Sweeting, 1995a, b). Example 4 in Section 2.4 was one illustration. The following example, which will also be used later for other purposes, is another standard example.

EXAMPLE 5. Suppose that, independently for $i = 1, \ldots, p$, $X_i \sim N(\mu_i, 1)$. The parameter of interest is

$$\theta = \frac{1}{p}|\boldsymbol{\mu}|^2 = \frac{1}{p}\sum_{i=1}^{p}\mu_i^2.$$

The usual choice of nuisance parameter here is $\lambda = \boldsymbol{\mu}/|\boldsymbol{\mu}|$, that is, the "direction" of $\boldsymbol{\mu}$ in R^p. Note that λ can also be viewed as a point on the surface of the unit ball in R^p; hence it is natural to assign λ the uniform prior on this surface. The resulting uniform-integrated likelihood (see Chang and Eaves, 1990 or Berger, Philippe and Robert,

1998) is

$$L^U(\theta) \propto \theta^{-(p-2)/4} \exp(-p\theta/2) \cdot I_{(p-2)/2}(\sqrt{p\theta}|x|), \tag{19}$$

where $|x| = (\sum_{i=1}^p x_i^2)^{1/2}$ and I_ν is the modified Bessel function of the first type of order ν.

For large p and assuming that θ stays bounded as $p \to \infty$, it can be shown that the mode of (19) is approximately

$$\hat{\theta} \approx \frac{1}{p}|x|^2 - \frac{(p-1)}{p} - \frac{1}{p}\left(\frac{2|x|^2}{(p-2)} - 1\right)^{-1}. \tag{20}$$

This is a sensible estimate. For instance, since $(|x|^2/p - E|X|^2/p) \to 0$ as $p \to \infty$ by the law of large numbers, it is immediate that $\hat{\theta} - \theta \to 0$. Thus $\hat{\theta}$ is consistent for θ.

In contrast, the profile likelihood does not yield a consistent estimator. Indeed, the profile likelihood for this problem can easily be seen to be

$$\hat{L}(\theta) \propto \exp(-(|x| - \sqrt{p\theta})^2), \tag{21}$$

which has mode $\hat{\theta} = |x|^2/p$. Clearly $\hat{\theta} - \theta \to 1$, establishing the inconsistency. (There do exist "classical" methods which would yield reasonable answers. One such is to look at the marginal distribution of $\sum_{i=1}^p X_i^2$, which is a noncentral chi-square distribution with parameter $p\theta$; the resulting likelihood, though complicated, will behave reasonably.)

3.2 Use in Bayesian Analysis

It is natural to seek to use an integrated likelihood, $L(\theta)$, via the Bayesian approach of choosing a prior distribution, $\pi(\theta)$, for the parameter of interest and obtaining the posterior distribution

$$\pi(\theta \mid x) \propto L(\theta)\pi(\theta).$$

This is completely legitimate when $\pi(\lambda|\theta)$ is a proper distribution, as observed in Section 1.3.2. When $\pi(\lambda|\theta)$ is improper, however, certain incoherencies such as marginalization paradoxes (Dawid, Stone and Zidek, 1973) can creep in, making this practice questionable. Fortunately, the practical effect of such incoherencies appears to be minor, and they can be minimized by appropriate choice of $\pi(\lambda|\theta)$ [and $\pi(\theta)$], such as the "reference prior" (see Section 4.4).

In Section 1.3.3, we listed some of the ways in which $L(\theta)$ can be of direct value to a Bayesian. There is an additional, rather subtle but important, reason for Bayesians to consider approaching the problem through integrated likelihood: one is considerably less likely to make a damaging mistake through use of improper prior distributions.

EXAMPLE 5 (Continued). Bayesians need not classify parameters as interesting or nuisance. Indeed, the common "naive" default Bayesian approach to this problem would be to base inference on the noninformative prior $\pi(\boldsymbol{\mu}) = 1$ for $\boldsymbol{\mu} = (\mu, \ldots, \mu_p)$; the resulting posterior distribution, given $\mathbf{x} = (x_1, \ldots, x_p)$, is

$$\pi(\boldsymbol{\mu} \mid \mathbf{x}) = (2\pi)^{-p/2} \exp\left\{-\sum_{i=1}^p (\mu_i - x_i)^2/2\right\}. \tag{22}$$

If, now, one is interested in $\theta = (\sum_{i=1}^p \mu_i^2)/p$, one can easily determine the posterior distribution $\pi(\theta|\mathbf{x})$ for θ, since, from (22), $p\theta$ has a noncentral chi-square distribution with p degrees of freedom and noncentrality parameter $|\mathbf{x}|^2 = \sum_{i=1}^p x_i^2$.

This posterior for θ is "bad." For instance, a common estimate of θ is the posterior mean, which here would be $\hat{\theta} = |\mathbf{x}|^2/p + 1$. This is badly inconsistent as $p \to \infty$, in the sense that $\hat{\theta} - \theta \to 2$. The posterior median and posterior mode also exhibit this behavior.

Thinking in terms of θ and the nuisance parameters λ in Section 3.1 avoided this problem. It was "obvious" to use the uniform distribution on λ to integrate out the nuisance parameters, and the ensuing integrated likelihood in (19) will work well with any sensible $\pi(\theta)$. The "error" in use of $\pi(\boldsymbol{\mu}) = 1$ above can be seen by changing variables to (θ, λ). Then $\pi(\boldsymbol{\mu}) = 1$ is transformed into

$$\pi(\theta, \lambda) = \theta^{-(p-2)/2}\, \pi(\lambda \mid \theta),$$

where $\pi(\lambda|\theta)$ is the uniform distribution on the surface of the unit ball, as before. Thus, $\pi(\boldsymbol{\mu}) = 1$ has unwittingly introduced a drastic and unreasonable prior distribution on θ. Separately considering λ and θ, through integrated likelihood, can help to avoid this type of error. (It should be noted that sophisticated default Bayesian approaches, such as the reference prior approach, automatically avoid this type of error; hence formal consideration of integrated likelihood is not strictly necessary if reference priors are employed.)

We have not discussed which priors $\pi(\theta)$ should be used with $L(\theta)$. Subjective choices, when available, are to be encouraged. Default choices and examples of the possible importance of the choice are discussed in Section 5.1.

3.3 Use in Empirical Bayes Analysis

In a variety of situations, including empirical Bayes analysis, it is common to estimate nuisance parameters using integrated likelihood and then to replace them in the likelihood with their estimates. This also goes under the name "Type II maximum

likelihood" (see Good, 1983). We content ourselves here with an example.

EXAMPLE 2 (Continued). Suppose now that $\theta = \boldsymbol{\mu} = (\mu, \ldots, \mu_p)$ is of interest, with nuisance parameter $\lambda = (\xi, \tau^2)$. The joint "likelihood" of all parameters is

$$L(\theta, \lambda) \propto \exp\left\{ -\sum_{i=1}^{p} (x_i - \mu_i)^2/2 \right\} \cdot \exp\left\{ -\sum_{i=1}^{p} (\mu_i - \xi)^2/(2\tau^2) \right\}, \tag{23}$$

which already includes the given prior distribution, $\pi(\boldsymbol{\mu}|\xi, \tau^2)$. In any case, to eliminate ξ and τ^2 the standard method is to form the integrated likelihood from (7),

$$L(\lambda) = L(\xi, \tau^2) = \int L(\theta, \lambda)\, d\theta \propto (1+\tau^2)^{-p/2} \exp\left\{ \frac{-p[s^2 + (\overline{x} - \xi)^2]}{2(1+\tau^2)} \right\},$$

estimate (ξ, τ^2) by the mode

$$\hat{\xi} = \overline{x}, \qquad \hat{\tau}^2 = \max\{0, s^2 - 1\}$$

and plug back into (23). After simplification, the result is

$$L(\theta) \propto \exp\left\{ -\frac{1}{2v} \sum_{i=1}^{p} (\mu_i - m_i)^2 \right\}, \tag{24}$$

where $v = \hat{\tau}^2/(1 + \hat{\tau}^2)$ and $m_i = vx_i + (1 - v)\overline{x}$. This is actually typically interpreted directly as the posterior distribution of $\boldsymbol{\mu}$.

Bayesians would argue that a superior approach is to integrate directly,

$$L(\theta) = \int L(\theta, \lambda)\pi(\lambda)\, d\lambda \tag{25}$$

[note that it would be incorrect to use the conditional prior $\pi(\lambda|\theta)$ here, since $L(\theta, \lambda)$ already contains $\pi(\theta|\lambda)$ as the second factor in (23); only the marginal prior for λ is still needed]. A common choice for $\pi(\lambda)$ is $\pi(\lambda) = \pi(\xi, \tau^2) = 1$ [although Berger and Strawderman, 1996, suggest that $\pi(\xi, \tau^2) = (1+\tau^2)^{-1}$ is better]. In a sense, the superiority of (25) over (24) is almost obvious, since (e.g.) if $\hat{\tau}^2 = 0$, then $v = 0$ and $m_i = \overline{x}$, so that (24) would imply that all $\mu_i = \overline{x}$ with absolute certainty. This is another example of the potential inadequacy of failing to incorporate the uncertainty in nuisance parameters.

Although the direct integrated likelihood in (25) is arguably superior to (24), it is worth noting that estimation of λ in $L(\theta, \lambda)$ by Type II MLE is at least better than using profile likelihood. Indeed, the profile likelihood can be easily seen to be

$$\hat{L}(\theta) \propto \left(\sum_{i=1}^{p} (\mu_i - \overline{\mu})^2 \right)^{-p/2} \cdot \exp\left\{ -\sum_{i=1}^{p} (\mu_i - x_i)^2/2 \right\}. \tag{26}$$

It is hard to know what to do with $\hat{L}(\theta)$; the first term has a singularity along the line $\boldsymbol{\mu} = (c, c, \ldots, c)$, and it is far from clear which, if any, of the other modes are reasonable.

4. VERSIONS OF INTEGRATED LIKELIHOOD

As mentioned in the introduction, any conditional prior density $\pi(\lambda|\theta)$ can be used to define an integrated likelihood. In this section, we review the more common choices of default or noninformative conditional priors and discuss their strengths and weaknesses.

4.1 Proper Conditional Priors

When $\pi(\lambda|\theta)$ is naturally a proper conditional density, certain of the concerns with interpreting and using the resulting integrated likelihood disappear. In particular, the resulting integrated likelihood can unambiguously be interpreted as a likelihood and can be combined with any (proper) prior density for θ to produce a true posterior. Example 5 provides an illustration of this, with $\pi(\lambda|\theta)$—the uniform distribution on the sphere of rotations—clearly being proper. Example 4 offers another illustration, with both the Jeffreys and the uniform distributions proper on the unit interval.

Today it is popular, for computational reasons, to use "vague" proper priors (often "vague" conjugate priors). Unfortunately, the use of such does not really provide protection against the concerns that arise with use of improper conditional priors; if a difficulty would arise in using an improper conditional prior, the same difficulty would manifest itself through great sensitivity to the degree of "vagueness" chosen.

4.2 Uniform Conditional Prior

The uniform choice of $\pi(\lambda|\theta) = 1$ has already been mentioned in the introduction and in several of the examples. It is still the most commonly used default conditional prior and is an attractive choice when nothing else is available. (When λ is a vector, many even prefer the uniform prior to the Jeffreys prior because of concerns with the behavior of the Jeffreys prior in higher dimensions.) There are,

however, well-documented difficulties with the uniform prior, perhaps the most well known being its lack of invariance to reparameterization. (Using the uniform prior for a positive parameter λ will yield a different integrated likelihood than using the uniform prior for its logarithm $\lambda^* = \log\lambda$, for example.) While this tends not to be a serious issue in practice (Laplace, 1812, suggested that parameterizations are typically chosen so that a uniform prior is reasonable), it does suggest that a uniform prior cannot be the final answer.

A more serious potential problem with the uniform conditional prior is that the resulting integrated likelihood may not exist. Here is a simple example.

EXAMPLE 6. Suppose X_1 and X_2 are independent $N(\theta, \sigma^2)$ and the uniform prior $\pi(\sigma^2|\theta) = 1$ is used for the nuisance parameter σ^2. Then

$$L^U(\theta) = \int_0^\infty \frac{1}{2\pi\sigma^2} \cdot \exp\left\{-\frac{1}{2\sigma^2}[(x_1-\theta)^2 + (x_2-\theta)^2]\right\} d\sigma^2 = \infty.$$

In contrast, use of the usual conditional prior $\pi(\sigma^2|\theta) = 1/\sigma^2$ (see Section 4.4) would yield

$$L^R(\theta) = \frac{1}{\pi[(x_1-\theta)^2 + (x_2-\theta)^2]}. \tag{27}$$

Interestingly, this latter integrated likelihood coincides with the profile likelihood and with the marginal likelihood.

4.3 Right Haar Measure

If $f(x|\theta, \lambda)$ is invariant with respect to an amenable group whose action on the parameter space leaves θ unchanged, then the compelling choice for $\pi(\lambda|\theta)$ is the induced right invariant Haar density for λ (see Berger, 1985, and Eaton, 1989, for definitions). Virtually all default Bayesian methods recommend this conditional prior, as do various "structural" and even frequentist approaches.

EXAMPLE 1 (Continued). The model is invariant under a location shift, and the right invariant Haar density for θ is the uniform density $\pi(\theta|\sigma^2) = 1$.

EXAMPLE 5 (Continued). The model is invariant under rotations of $\mathbf{x}$ and $\boldsymbol{\mu}$, and the rotation group leaves $\theta = |\boldsymbol{\mu}|^2/p$ unchanged. The right invariant Haar density (actually the Haar density here, since the rotation group is compact and hence unimodular) induced on $\lambda = \boldsymbol{\mu}/|\boldsymbol{\mu}|$ is the uniform density on the unit sphere discussed in Section 3.1.

For formal discussion and other examples of use of the right Haar density as the conditional prior, see Chang and Eaves (1990) and Datta and Ghosh (1995a). The chief limitation of this approach is the rarity of suitable invariance. A secondary limitation is that, if the group is not amenable, use of the resulting right Haar density can be problematical.

4.4 Conditional Reference Integrated Likelihood

The reference prior algorithm (Bernardo, 1979; Berger and Bernardo, 1989, 1992; Bernardo and Smith, 1994) is a quite general and powerful tool for obtaining "automatic" priors to be used in Bayesian analysis. It is motivated by trying to find that prior distribution which is least informative, in the sense of maximizing (in an asymptotic sense) the expected Kullback–Liebler divergence between the prior distribution and the posterior distribution. (Intuitively, such a prior distribution allows the data to "speak most loudly.") The reference prior is typically the same as the Jeffreys prior in the one-dimensional case; when the parameter space is multivariate, the reference algorithm takes into account the order of inferential importance of the parameters, by partitioning the parameter vector $\omega = (\omega_1, \omega_2, \ldots, \omega_p)$ into several blocks of decreasing interest to the investigator. Berger and Bernardo suggest using the one-at-time reference prior which corresponds to partitioning ω into p one-dimensional blocks.

Since the reference prior depends on which parameters are of primary interest, it is usually different from the Jeffreys prior. In numerous multivariate examples, it has been shown to perform considerably better than the Jeffreys prior. Also, it seems to typically yield procedures with excellent frequentist properties (Ghosh and Mukerjee, 1992; Liseo, 1993; Sun, 1994; Datta and Ghosh, 1995a, b).

In the standard Bayesian approach, the reference prior is used to produce the joint posterior distribution for (θ, λ); then λ is integrated out to obtain the marginal posterior for θ. In this approach, there is no need to develop the notion of a likelihood for θ. Indeed, any attempt to do so directly, via an expression such as (10), is made difficult by the typical impropriety of reference priors; one cannot define conditional and marginal distributions directly from an improper distribution.

Therefore, to produce a reference-integrated likelihood, we need to slightly modify the reference prior algorithm. In this paper, we only discuss the "two groups" case, where the parameter vector is split in the parameters of interest, θ, and the nuisance parameters, λ; extensions to several groups is immediate. For simplicity of notation, we will take θ and λ to be scalars.

Under certain regularity conditions (see, e.g., Bernardo and Smith, 1994), basically the existence of a consistent and asymptotically normal estimator of the parameters, the reference prior for λ, when θ is known, is defined to be

$$\pi^*(\lambda \mid \theta) \propto \sqrt{I_{22}(\theta, \lambda)}, \tag{28}$$

where $I_{22}(\theta, \lambda)$ is the lower right corner of the expected Fisher information matrix. Since direct use of this, as in (10), to obtain an integrated likelihood is problematical due to its typical impropriety (Basu, 1977), we employ the idea of the Berger and Bernardo (1992) reference prior algorithm and consider a sequence of nested subsets of $\Omega = \Theta \times \Lambda$, $\Omega_1, \Omega_2, \ldots$, increasing to Ω and over which (28) can be normalized; then the conditional reference prior, and associated integrated likelihood, will be defined as the appropriate limit. Indeed, defining $\Lambda_m(\theta) = \{\lambda: (\theta, \lambda) \in \Omega_m\}$ and the normalizing constant over this compact set as $K_m(\theta)^{-1} = \int_{\Lambda_m(\theta)} \pi^*(\lambda \mid \theta)\, d\lambda$, the *conditional reference prior* is defined as

$$\pi^R(\lambda \mid \theta) = h(\theta)\pi^*(\lambda \mid \theta), \tag{29}$$

where

$$h(\theta) = \lim_{m \to \infty} \frac{K_m(\theta)}{K_m(\theta_0)}, \tag{30}$$

assuming the limit is unique up to a proportionality constant for any θ_0 in the interior of Θ. The corresponding integrated likelihood is then given by

$$L^R(\theta) = \int_\Lambda f(x \mid \theta, \lambda)\pi^R(\lambda \mid \theta)\, d\lambda, \tag{31}$$

which is in the "standard" form of an integrated likelihood.

This definition of the conditional reference prior can be shown to be consistent with the definition of the joint reference prior $\pi^R(\theta, \lambda)$ in Berger and Bernardo (1992), providing the joint reference prior exists and (30) holds, in the sense that then

$$\pi^R(\theta, \lambda) = \pi^R(\lambda \mid \theta)\pi^R(\theta) \tag{32}$$

for some function $\pi^R(\theta)$. We will then define $\pi^R(\theta)$ to be the *marginal reference prior* for θ. The conditional reference prior can be shown to share many of the desirable properties of the joint reference prior, such as invariance to reparameterization of the nuisance parameters. Note that this conditional reference prior was also considered in Sun and Berger (1998), although for different purposes.

EXAMPLE 7. *The coefficient of variation.* Let $X_1, X_2, \ldots, X_n$ be n iid random variables with distribution $N(\mu, \sigma^2)$. The parameter of interest is the coefficient of variation $\theta = \sigma/\mu$ and $\lambda = \sigma$ is the nuisance parameter. The expected Fisher information matrix, in the (θ, λ) parameterization, is

$$I(\theta, \lambda) = \begin{pmatrix} \frac{1}{\theta^4} & -\frac{1}{\lambda\theta^3} \\ -\frac{1}{\lambda\theta^3} & \frac{2\theta^2+1}{\lambda^2\theta^2} \end{pmatrix}. \tag{33}$$

The above algorithm gives

$$\pi^*(\lambda \mid \theta) \propto \frac{1}{\lambda}\sqrt{\frac{2\theta^2+1}{\theta^2}}. \tag{34}$$

A natural sequence of compact sets in the (μ, σ) parameterization is given by

$$\Omega_m = \left\{(\mu, \sigma): -a_m < \mu < a_m, \ \frac{1}{b_m} < \sigma < b_m\right\}$$

for increasing sequences a_m and b_m that diverge to infinity. Then the resulting $\Lambda_m(\theta)$ sequence is

$$\Lambda_m(\theta) = \begin{cases} (1/b_m, b_m), & |\theta| > \dfrac{b_m}{a_m}, \\ (1/b_m, |\theta|a_m), & \dfrac{1}{a_m b_m} < |\theta| < \dfrac{b_m}{a_m}, \\ \varnothing, & |\theta| < \dfrac{1}{a_m b_m}. \end{cases}$$

Therefore

$$K_m^{-1} = \begin{cases} 2\sqrt{\dfrac{2\theta^2+1}{\theta^2}}\log b_m, & |\theta| > \dfrac{b_m}{a_m}, \\ \sqrt{\dfrac{2\theta^2+1}{\theta^2}}\log(|\theta|a_m b_m), & \dfrac{1}{a_m b_m} < |\theta| < \dfrac{b_m}{a_m}, \\ 0, & |\theta| < \dfrac{1}{a_m b_m}, \end{cases}$$

and

$$h(\theta) = \lim_{m \to \infty} \frac{K_m(\theta)}{K_m(\theta_0)} \propto \sqrt{\frac{\theta^2}{2\theta^2+1}}.$$

Thus the conditional reference prior and integrated likelihood are $\pi^R(\lambda \mid \theta) = 1/\lambda$ and

$$\begin{aligned} L^R(\theta) &= \int_0^\infty \frac{1}{\lambda} f(x \mid \theta, \lambda)\, d\lambda \\ &\propto \exp\left[-\frac{n}{2\theta^2}\left(1 - \frac{\overline{x}^2}{D^2}\right)\right]\int_0^\infty z^{n-1} \\ &\quad \cdot \exp\left[-\frac{n}{2}D^2\left(z - \frac{\overline{x}}{D^2\theta}\right)^2\right] dz, \end{aligned} \tag{35}$$

where $\overline{x}$ is the sample mean and $D^2 = \sum x_i^2/n$. [Note, also, that the problem is invariant to scale

changes, and $\pi(\lambda \mid \theta) = 1/\lambda$ is the resulting right invariant Haar density.] This example will be further discussed in Section 5.1.

EXAMPLE 5 (Continued). When $\theta = (\sum_1^p \mu_i^2)/p$ is the parameter of interest, then the conditional reference prior for the nuisance parameters λ is simply the uniform prior on the surface of the unit ball, as discussed in Section 3.1.

EXAMPLE 6 (Continued). Unlike the conditional uniform prior, the conditional reference prior for this model, namely $\pi^R(\lambda \mid \theta) \propto 1/\lambda$, yielded a finite integrated likelihood in (27). Of course, with three or more observations, $L^U(\theta)$ would also be finite here, but the example is indicative of the commonly observed phenomenon that reference priors virtually always yield finite integrated likelihoods, while the uniform prior may not.

4.5 Other Integrated Likelihoods

In one sense, there are as many integrated likelihoods as there are priors. For instance, any method which yields a noninformative prior, $\pi^N(\theta, \lambda)$, leads to an integrated "likelihood"

$$L(\theta) \propto \int L(\theta, \lambda)\, \pi^N(\theta, \lambda)\, d\lambda. \tag{36}$$

Since $\pi^N(\theta, \lambda)$ is a joint prior distribution on both θ and λ, however, this is not formally in the spirit of (3); equation (36) would actually yield the proposed marginal posterior distribution for θ and must be normalized by dividing by the marginal $\pi^N(\theta) = \int \pi^N(\theta, \lambda)\, d\lambda$ (if finite) to yield a likelihood.

Reference noninformative priors offer a ready solution to this dilemma, since the reference prior algorithm itself suggested a suitable $\pi^R(\lambda|\theta)$ for use in (3). While it is easy to normalize proper priors correctly, other noninformative priors are more difficult to modify to produce an integrated likelihood. We illustrate the possibility by considering how to modify the Jeffreys prior approach to yield an integrated likelihood, since the Jeffreys prior is probably the most widely used default prior distribution. (The motivation for the Jeffreys prior can be found in Jeffreys, 1961, and was primarily the desire to construct default priors such that the resulting Bayesian answers are invariant under reparameterization.)

The Jeffreys prior for (θ, λ) is

$$\pi^J(\theta, \lambda) \propto \sqrt{\det(I(\theta, \lambda))},$$

where $I(\theta, \lambda)$ is the expected Fisher information matrix. To obtain a reasonable $\pi^J(\lambda|\theta)$, perhaps the most natural option is to simply treat θ as given, and derive the Jeffreys prior for λ with θ given. It is easy to see that the result would be

$$\pi^*(\lambda \mid \theta) \propto \sqrt{\det(I_{22}(\theta, \lambda))}, \tag{37}$$

where $I_{22}(\theta, \lambda)$ is the corner of $I(\theta, \lambda)$ corresponding to the information about λ. This, however, also has ambiguities. In Example 7, for instance, we found from (34) that $\pi^*(\lambda|\theta) \propto \lambda^{-1}\sqrt{2 + \theta^{-2}}$ but, since θ is now to be viewed as given, the factor $\sqrt{2 + \theta^{-2}}$ is just a proportionality constant which can be ignored (the problem does not arise for proper priors, since such a factor would disappear in the normalizing process).

The net result of this reasoning suggests that the correct definition of $\pi^J(\lambda|\theta)$ is as given in (37), but ignoring any multiplicative factors which only involve θ. Thus, in Example 7, we would obtain $\pi^J(\lambda|\theta) = 1/\lambda$ which is the same as $\pi^R(\lambda|\theta)$. We have not explored the quality of integrated likelihoods based on $\pi^J(\lambda|\theta)$, but suspect that they would typically be satisfactory.

The other prominently studied default priors are the *probability matching priors*, designed to produce Bayesian credible sets which are optimal frequentist confidence sets in a certain asymptotic sense. Literature concerning these priors can be accessed through the recent papers by Berger, Philippe and Robert (1998), Datta and Ghosh (1995a, b) and Ghosh and Mukerjee (1992). We have not explored the use of these priors in defining integrated likelihood.

5. LIMITATIONS OF INTEGRATED LIKELIHOOD

5.1 The Posterior Distribution May Be Needed

As in the univariate case without nuisance parameters, the integrated likelihood function contains the information provided by the data (filtered by the conditional prior on λ) and often can be used directly for inferential purposes. In some cases, however, $L^*(\theta)$ needs to be augmented by a prior distribution for θ (possibly noninformative), yielding a posterior distribution for θ, before it can serve as a basis for inference. (We are envisaging that the posterior distribution would be used in ordinary Bayesian ways, to construct error estimates, credible sets, etc.)

EXAMPLE 4 (Continued). In Section 2.4, we used only the modes of $L^U(N)$ and $L^J(N)$ as point estimates for N. How can we do more, for example, convey the precision of the estimates?

Classical approaches have not made much headway with this problem, and the "standard" noninformative prior Bayesian approach also fails.

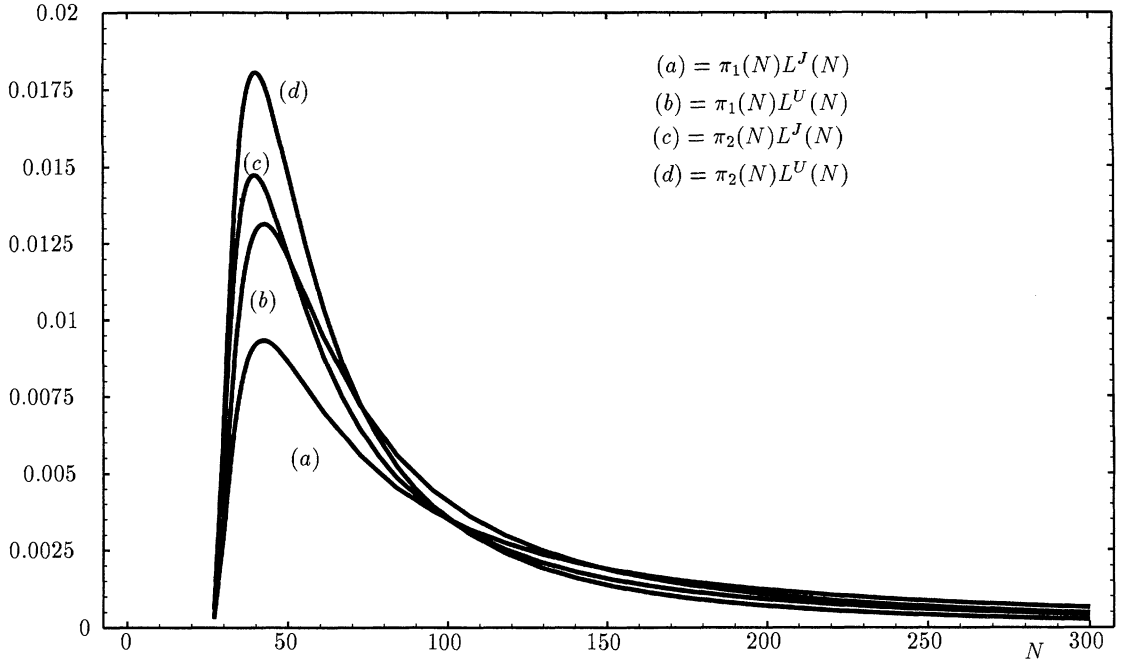

FIG. 4. *Posterior distributions proportional to* $L^U(N)\pi_1(N)$, $L^J(N)\pi_1(N)$, $L^U(N)\pi_2(N)$ *and* $L^J(N)\pi_2(N)$.

The standard noninformative prior for N would be $\pi(N) = 1$. But $L^a(N)$ behaves like cN^{-a} for large N (see Kahn, 1987), as does the resulting posterior, so for $a \leq 1$ the posterior will be improper and Bayesian analysis cannot succeed. (In a sense the difficulty here is that sophisticated noninformative prior methodology does not really exist for discrete N; see the next example for an illustration of what such methodology can do for a continuous parameter.)

Subjective Bayesians would argue the importance of introducing proper subjective prior distributions here, and it is hard to disagree. This could be done either by keeping the (vague) $\pi(N) = 1$, but introducing a Beta (a, b) prior for θ with $a > 1$ (in which case Kahn, 1987, can be used to show that the posterior is proper), or by choosing an appropriately decreasing $\pi(N)$.

One could still, however, argue for the benefits of having a default or conventional analysis available for the problem. To go along with the default $L^U(N)$ or $L^J(N)$ [see (16) and (17)], one might consider, as default priors, either $\pi_1(N) = 1/N$ (Raftery, 1988; Moreno and Girón, 1995; de Alba and Mendoza, 1996) or the Rissanen (1983) prior

$$\pi_2(N) \propto \prod_{i=0}^{k_N} (\log^{(i)}(N))^{-1},$$

where $\log^{(0)}(N) = N$ and $\log^{(k+1)}(N) = \log\log^{(i)}(N)$, with k_N the largest integer such that $\log^{(k_N)}(N) > 1$. This latter prior is, in some sense, the vaguest possible proper prior. Both π_1 and π_2 can easily be shown to yield proper posteriors when paired with either $L^U(N)$ or $L^J(N)$. Figure 4 shows the four resulting posterior distributions for the data $\mathbf{s} = (16, 18, 22, 25, 27)$.

Inferential conclusions are straightforward from these proper posteriors. For instance, the quartiles of the four posteriors are given in Table 2. (Providing quartiles is more reasonable than providing moments when, as here, the distributions have large and uncertain tails.)

Compare this with the much more limited (and questionable) inferences provided in Table 1. While the answers in Table 2 vary enough that a subjective Bayesian analysis might be deemed by many to be necessary, a case could also be made for choosing, as a conventional analysis, the $L^U\pi_1$ or $L^J\pi_2$ posterior distributions.

The Gleser–Hwang class. Example 4 is a rather special situation, because the parameter is discrete. However, similar phenomena occur for a large

TABLE 2
Quartiles of the posterior distributions for N

Posterior	First quartile	Median	Third quartile
$L^U(N)\ \pi_1(N)$	51	80	158
$L^J(N)\ \pi_1(N)$	59	112	288
$L^U(N)\ \pi_2(N)$	44	63	106
$L^J(N)\ \pi_2(N)$	47	74	152

class of problems, which includes Fieller's problem, errors-in-variables models, and calibration models. We begin discussion of this class by presenting an example and then discuss the characterization of the class given by Gleser and Hwang (1987).

EXAMPLE 7 (Continued). The use of the conditional reference prior, $\pi(\lambda|\theta) \propto 1/\lambda$, for the nuisance parameter $\lambda = \sigma$ in the coefficient of variation problem leads to the integrated likelihood (35). It can be easily seen that this likelihood does not go to zero as $|\theta|$ goes to infinity, making direct inferential use of $L^R(\theta)$ difficult. To progress, one can bring in the marginal reference prior

$$\pi^R(\theta) \propto \frac{1}{|\theta|\sqrt{\theta^2 + 1/2}}. \tag{38}$$

This leads to the reference posterior distribution,

$$\begin{aligned}\pi^R(\theta \mid x) &\propto \frac{1}{|\theta|\sqrt{\theta^2+1/2}} \exp\left[-\frac{n}{2\theta^2}\left(1-\frac{\overline{x}^2}{D^2}\right)\right] \\ &\quad \cdot \int_0^\infty z^{n-1} \exp\left[-\frac{n}{2}D^2\left(z-\frac{\overline{x}}{D^2\theta}\right)^2\right] dz,\end{aligned} \tag{39}$$

which, interestingly, is proper, and hence can be used for inference. This is an example of the rather amazing property of reference (and Jeffreys) priors that they virtually always seem to yield proper posterior distributions, at least in the continuous parameter case.

Of course, one might object to using (39) for inference, either because a Bayesian analysis is not wanted or because the introduction of the reference prior seems rather arbitrary. Indeed, it is difficult to come up with auxiliary supporting arguments for use of the reference prior here. For instance, the resulting Bayesian inferences will often not have particularly good frequentist properties, which is one of the commonly used arguments in support of reference priors.

It is important, however, to place matters in perspective. For this situation, there simply are no methods of "objective" inference that will be viewed as broadly satisfactory. Consider likelihood methods, for instance. The profile likelihood for θ can be shown to be

$$\hat{L}(\theta) \propto \left(\frac{\theta}{g(x,\theta)}\right)^n \cdot \exp\left[-\frac{n}{2\theta^2} + \frac{2n}{g^2(x,\theta)}(\overline{x}g(x,\theta) - \theta^2 D^2)\right],$$

where $g(x,\theta) = -\overline{x} + (\operatorname{sgn}\theta)\sqrt{\overline{x}^2 + 4D^2\theta^2}$. It is easy to see that, as $|\theta| \to \infty$, this approaches a positive constant. The correction factor for the profile likelihood given by the Cox–Reid (1987) method (see the Appendix) yields the conditional profile likelihood

$$\begin{aligned}\hat{L}^C(\theta) &\propto \left\{|\theta|(\overline{x}\operatorname{sgn}\theta + \sqrt{\overline{x}^2 + 4D^2\theta^2})\right\} \\ &\quad \cdot \left\{\sqrt{\theta^2+1/2} \cdot \sqrt{4\theta^2D^4 + \overline{x}^2(\overline{x}\operatorname{sgn}\theta + \sqrt{\overline{x}^2+4D^2\theta^2})^2}\right\}^{-1} \\ &\quad \cdot \hat{L}(\theta).\end{aligned} \tag{40}$$

It is easy to see that this, too, is asymptotically constant.

The similarities between the three different likelihoods [$L^R(\theta)$, $\hat{L}(\theta)$ and $\hat{L}^C(\theta)$] are better appreciated in the special case where $\overline{x} = 0$ and $D = 1$. Then

$$\begin{aligned}L^R(\theta) &\propto \hat{L}(\theta) \\ &\propto \frac{\sqrt{\theta^2+1/2}}{|\theta|}\hat{L}^C(\theta) \propto \exp\left(-\frac{n}{2\theta^2}\right).\end{aligned} \tag{41}$$

These are graphed in Figure 5, along with $\pi^R(\theta|x)$. Note that the conditional profile likelihood is virtually indistinguishable from the other likelihoods. There is no appealing way to use these likelihoods for inference.

Frequentists might argue that this simply reveals that no type of likelihood or Bayesian argument is appealing for this example. However, frequentist conclusions are also very problematical here, since this example falls within the class of problems, identified in Gleser and Hwang (1987), where frequentist confidence procedures must be infinite sets (and often the whole parameter space) with positive probability. [Saying that a 95% confidence set is $(-\infty, \infty)$ is not particularly appealing]. First, we restate the Gleser and Hwang result in our notation.

THEOREM 1 (Gleser and Hwang, 1987). *Consider a two parameter model with sampling density $f(x|\theta,\lambda)$, $\theta \in \Theta$, $\lambda \in \Lambda$, on the sample space $\mathcal{X}$. Suppose there exists a subset Θ^* of Θ and a value of λ^* in the closure of Λ such that θ has an unbounded range over Θ^* and such that for each fixed $\theta \in \Theta^*$ and $x \in \mathcal{X}$,*

$$\lim_{\lambda \to \lambda^*} f(x \mid \theta, \lambda) = f(x \mid \lambda^*) \tag{42}$$

exists, is a density on $\mathcal{X}$ and is independent of θ. Then every confidence procedure $C(X)$ for θ with positive confidence level $1-\alpha$ will yield infinite sets with positive probability.

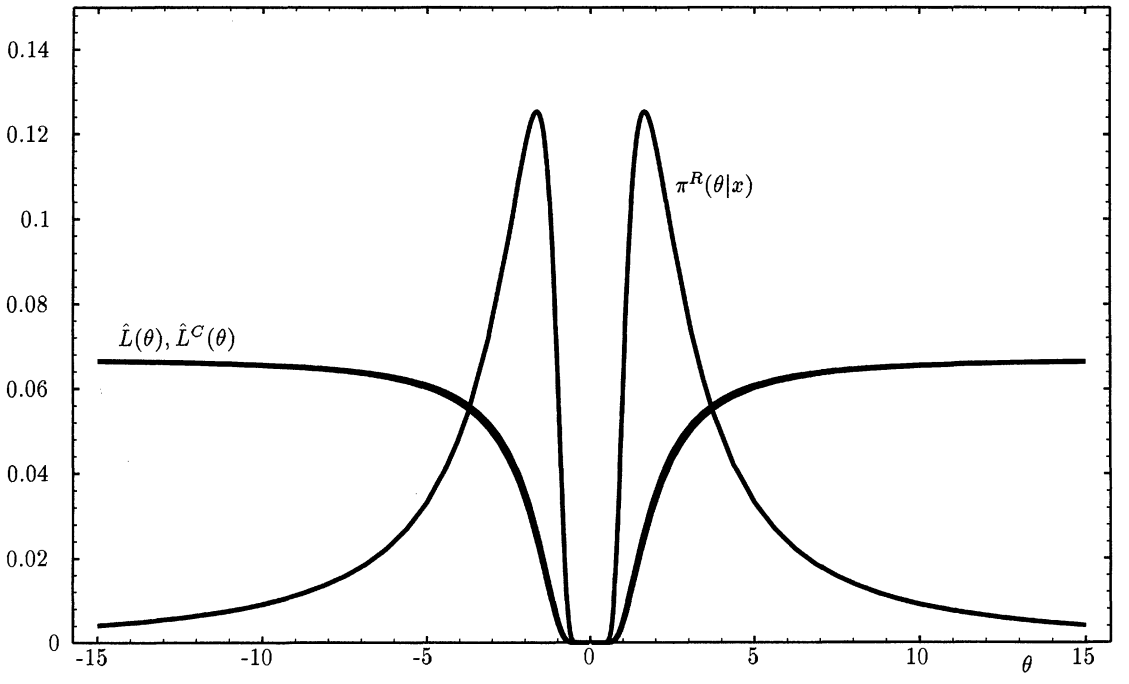

FIG. 5. *Reference, profile and conditional likelihoods and reference posterior for the coefficient of variation problem, with* $\bar{x} = 0$, $n = 5$ *and* $D = 1$.

To show that Example 7 is an example of this type, we must slightly generalize Theorem 1, as follows.

THEOREM 1′. *Assume there exist a sequence* (θ_m, λ_m) *and a value* λ^*, *with* $\lambda_m \to \lambda^*$ *and* $|\theta_m| \to \infty$, *such that*

$$\lim_{m\to\infty} f(x \mid \theta_m, \lambda_m) = f(x \mid \lambda^*), \tag{43}$$

for some density $f(x \mid \lambda^*)$. *Then the conclusion of Theorem* 1 *holds.*

The proof in Gleser and Hwang (1987) can be followed exactly, using the new condition in obvious places.

EXAMPLE 7 (Continued). If $\theta = \sigma/\mu$ and the nuisance parameter is chosen to be $\lambda = \mu$, then defining $\lambda_m = 1/|\theta_m|$ with $|\theta_m| \to \infty$ yields

$$\lim_{m\to\infty} f(x \mid \theta_m, \lambda_m) = \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{nD^2}{2}\right\},$$

so that (43) is satisfied. Hence frequentist confidence sets must be infinite with positive probability.

Another interesting fact is that the class of densities satisfying (43) appears to be related to the class of densities for which the profile likelihood does not go to zero at infinity.

LEMMA 1. *Under the conditions of Theorem* 1′, *the profile likelihood does not converge to zero.*

PROOF. From condition (43),

$$\lim_{m\to\infty} \sup_{\lambda\in\Lambda} f(x \mid \theta_m, \lambda) \geq \lim_{m\to\infty} f(x \mid \theta_m, \lambda_m) = f(x \mid \lambda^*) > 0.$$

Recall that this class of densities, which are very problematic from either a likelihood or a frequentist perspective, includes a number of important problems (in addition to the coefficient of variation problem) such as Fieller's problem, errors-in-variables models and calibration models. One can make a convincing argument that such problems are simply not amenable to any "objective" analysis; prior information is crucial and must be utilized. If, however, one is unwilling or unable to carry out a true subjective Bayesian analysis, then a case can be made for having standard "default" analyses available. The Bayesian analyses with reference noninformative priors, such as that leading to $\pi^R(\theta|x)$ in Example 7, are arguably the best candidates for such "default" analyses.

Visualization. We have argued that, for certain important classes of problems, inference appears impossible based on $L(\theta)$ alone. A counterargument is sometimes advanced, to the effect that formal inference is not necessary; it may suffice simply

to present $L(\theta)$ itself as the conclusion, with consumers learning to interpret likelihood functions directly.

One difficulty with this argument is that likelihood functions can appear very different, depending on the parameterization used.

EXAMPLE 7 (Continued). The conditional reference likelihood, $L^R(\theta)$, for $\overline{x} = 1$, $n = 2$ and $D = 2$, is presented in Figure 6a. Suppose, instead, that the parameterization $\xi = \theta/(1+|\theta|)$ had been used. The conditional reference likelihood for ξ can then easily be seen to be

$$\tilde{L}^R(\xi) = L^R(\xi/(1-|\xi|)),$$

which is plotted in Figure 6b over $(-1, 1)$, the range of ξ. Visually, $\tilde{L}^R(\xi)$ and $L^R(\theta)$ appear to convey markedly different information: for $\tilde{L}^R(\xi)$ the "local" mode appears to dominate, while for $L^R(\theta)$ it appears to be insignificant relative to the huge tail. Yet these two functions clearly reflect the same information.

Introducing $\pi(\theta)$ removes this difficulty [providing that the method used for obtaining $\pi(\theta)$ is invariant under reparameterization of θ], because the Jacobian will correct for the influence of the change of variables. Subjective priors will be suitably invariant, as are the Jeffreys and reference priors (see Datta and Ghosh, 1996).

EXAMPLE 7 (Continued). The marginal reference prior, $\pi^R(\theta)$ in (38), results in the proper posterior in (39). If, instead, the ξ parameterization had been used, the marginal reference prior for ξ would be

$$\tilde{\pi}^R(\xi) = \pi^R(\xi/(1-|\xi|)) \cdot (1-|\xi|)^{-2},$$

which is obtainable from $\pi^R(\theta)$ by straightforward change of variables. Because of this invariance of the prior, it is clear that the two posteriors, $\pi^R(\theta \mid x) \propto L^R(\theta)\pi^R(\theta)$ and $\tilde{\pi}(\xi \mid x) \propto \tilde{L}^R(\xi)\tilde{\pi}^R(\xi)$, are simple transformations of each other. Figure 7a and 7b graphs these two posteriors for the same situation as in Figure 6. These seem to be visually satisfactory, in the sense of conveying similar information.

One possibility for "correcting" likelihoods to prevent this visualization problem is to mimic the Bayesian approach by including a "Jacobian" term to account for reparameterization. Various suggestions to this effect have been put forth in the likelihood literature, from time to time, but none seem to have caught on.

5.2 Premature Elimination of Nuisance Parameters

Any summary reduction of data runs the risk of losing information that might later be needed, and nuisance parameter elimination is no exception. We review a few of the ways in which this loss of information can occur. The most basic situation that can cause problems is if more data are later to be observed.

EXAMPLE 1 (Continued). Defining $S_1^2 = \sum_{i=1}^n (x_i - \overline{x})^2/n$, the uniform integrated likelihood for σ^2 was $L_1^U(\sigma^2) \propto \sigma^{-(n-1)} \exp\{-nS_1^2/2\sigma^2\}$. Suppose additional data $x_{n+1}, \ldots, x_{n+m}$ become available. The corresponding uniform integrated likelihood for σ^2 from these data alone is $L_2^U(\sigma^2) \propto \sigma^{-(m-1)} \exp\{-mS_2^2/2\sigma^2\}$, with S_2^2 being the corresponding sum of square deviations. To find the overall likelihood for σ^2, it is tempting to multiply $L_1^U(\sigma^2)$ and $L_2^U(\sigma^2)$, since the two data sets were independent (given the parameters). Note, however, that, if all the data were available to us, we would use the integrated likelihood

$$L_3^U(\sigma^2) \propto \sigma^{-(n+m-1)} \exp\left\{\frac{-(n+m)S_3^2}{2\sigma^2}\right\},$$

where S_3^2 is the overall sum of square deviations, and this is not the product of L_1^U and L_2^U. Indeed, from knowledge only of L_1^U and the new data, L_3^U cannot be recovered. (To recover L_3^U, one would also need $\overline{x}$ from the original data.)

A second type of loss of information can occur in meta-analysis, where related studies are analyzed.

EXAMPLE 2 (Continued). We generalize this example slightly, supposing the observations to be $X_{ij} \sim N(\mu_i, \sigma_i^2)$, $j = 1, \ldots, n_i$ and $i = 1, \ldots, p$. Of interest here are the study-specific μ_i or the overall (ξ, τ^2), where $\mu_i \sim N(\xi, \tau^2)$, $i = 1, \ldots, p$. Such data might arise from studies at p different sites, with the means at the different sites being related to the overall population mean, ξ, as indicated. Each site might choose to eliminate the nuisance parameter σ_i^2 by integration [using, say, $\pi^R(\sigma_i^2) = 1/\sigma_i^2$], resulting in the report of the integrated likelihoods

$$L_i^R(\mu_i) \propto \left[S_i^2 + (\overline{x}_i - \mu_i)^2\right]^{-n_i/2}, \tag{44}$$

where $\overline{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$ and $S_i^2 = \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^2/n_i$, $i = 1, \ldots, p$. One might then be tempted to use the product of these likelihoods [along with the knowledge that $\mu_i \sim N(\xi, \tau^2)$] in the meta-analysis.

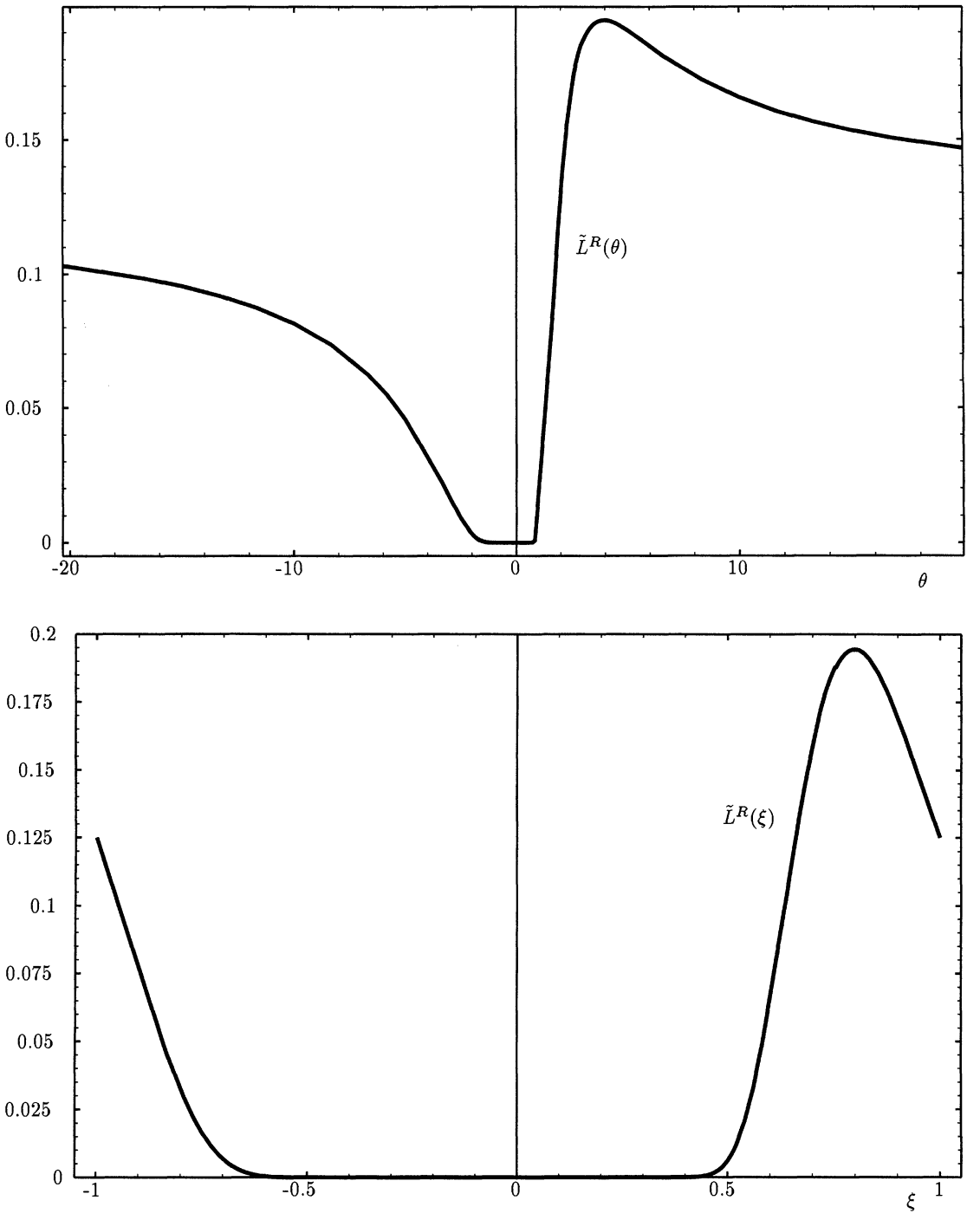

FIG. 6. *Integrated likelihoods for Example 7 when* $\overline{x} = 1$, $n = 2$ *and* $D = 2$. (a) θ *parameterization*; (b) ξ *parameterization.*

If the σ_i^2 were completely unrelated, such an analysis would be reasonable. Typically, however, the σ_i^2 would themselves be closely related, as in Hui and Berger (1983), and their "independent" elimination by integration would then not be appropriate. Interestingly, in this situation the original likelihood can be recovered (assuming the original model is known), in that the sufficient statistics n_i, $\overline{x}_i$ and S_i^2 can all be found from (44). But one would have to use these sufficient statistics to reconstruct the full likelihood and not use the $L_i^R(\mu_i)$ directly.

A third situation in which one should not use an integrated likelihood for θ is when prior informa-

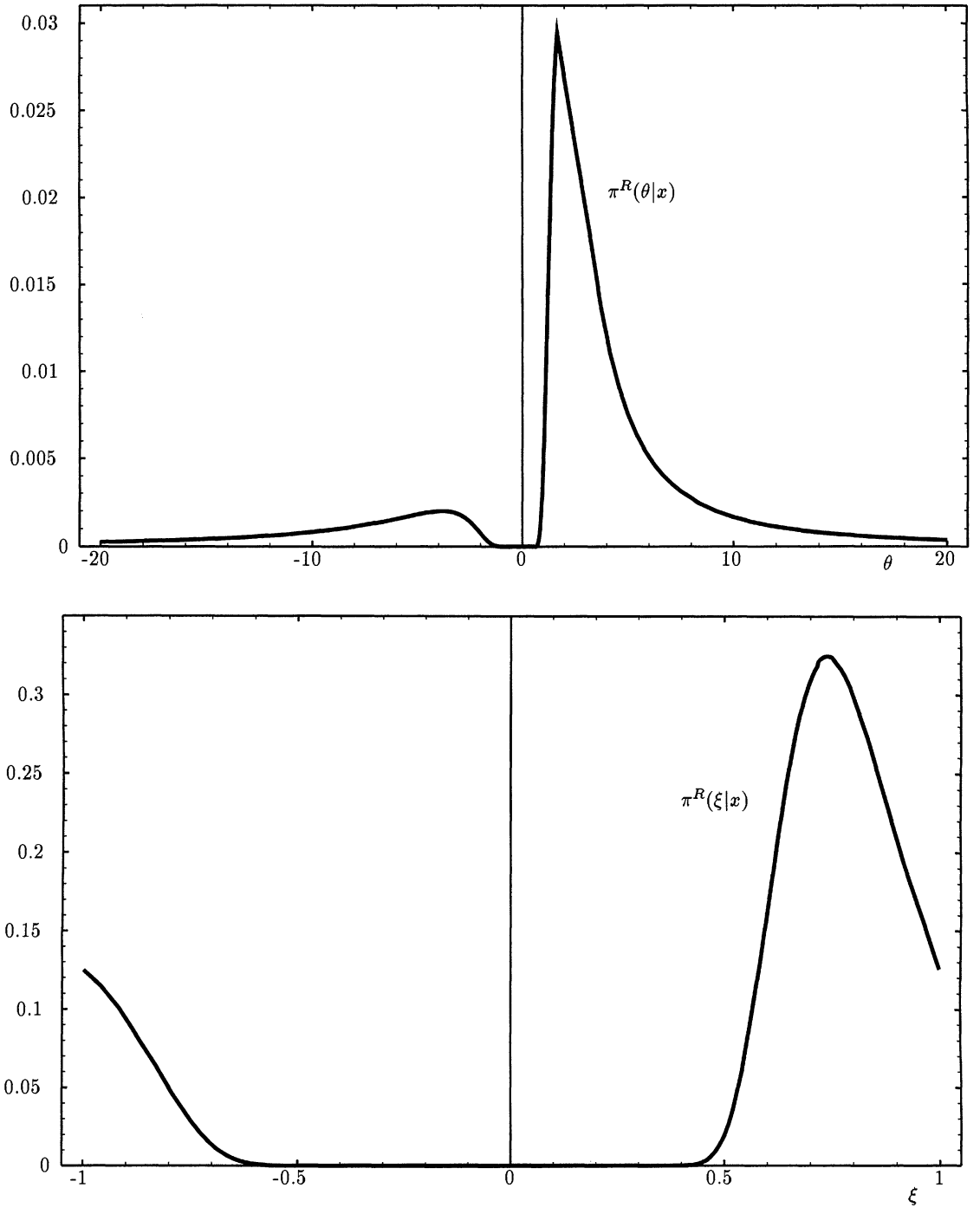

FIG. 7. *The reference posteriors for the situation of Figure* 6; (a) θ *parameterization*; (b) ξ *parameterization.*

tion is available in the form of a conditional prior distribution $\pi(\theta|\lambda)$. If λ is first eliminated from the likelihood, then one cannot subsequently make use of $\pi(\theta|\lambda)$. [One possible default treatment of this situation is described in Sun and Berger (1998): first find the reference marginal prior for λ, based on $\pi(\theta|\lambda)$, and then find the resulting posterior for θ.]

While one should be aware of the potential problems in premature elimination of nuisance parameters, the situation should be kept in perspective. Problems arise only if the nuisance parameters contain a significant amount of information about how the data relate to the quantities of interest; in such a situation one should delay integrating away the

nuisance parameters as long as possible. In the examples discussed above, however, the nuisance parameters hold little enough information about the quantities of interest that the answers obtained by simply multiplying the integrated likelihoods are reasonably close to the answers from the "correct" likelihood. Indeed, much of statistical practice can be viewed as formally or informally eliminating nuisance parameters at different stages of the analysis, then multiplying the resulting likelihoods. The trick is to recognize when this is a reasonable approximation and when it is not.

APPENDIX

To derive a conditional profile likelihood for the coefficient of variation in Example 7, one first needs to obtain an orthogonal parameterization. It can be shown that $\xi = (\sqrt{2\sigma^2 + \mu^2})^{-1}$ is orthogonal to θ. Cox and Reid (1987) defined the conditional profile likelihood as

$$\hat{L}^C(\theta) = \hat{L}(\theta)|j_{\xi,\xi}(\theta, \hat{\xi}_\theta)|^{-1/2},$$

where $|j_{\xi,\xi}(\theta, \hat{\xi}_\theta)|$ is the lower right corner of the observed Fisher Information matrix, and $\hat{\xi}_\theta$ is the conditional maximum likelihood estimate. Calculations show that

$$|j_{\xi,\xi}(\theta, \hat{\xi}_\theta)|^{-1/2} \propto \frac{|\theta|\left(\overline{x} + \sqrt{\overline{x}^2 + 4D^2\theta^2}\right)}{\sqrt{\theta^2 + 1/2}\sqrt{4D^4\theta^2 + \overline{x}^2\left(\overline{x} + \sqrt{\overline{x}^2 + 4D^2\theta^2}\right)^2}}.$$

Together with the fact that the profile likelihood is invariant with respect to the choice of the nuisance parameters, one obtains expression (40).

ACKNOWLEDGMENTS

Supported in part by NSF Grants DMS-93-03556, DMS-96-26829 and DMS-9802261, and the Italian Consiglio Nazionale delle Ricerche. The authors are also grateful to referees and Jan Bjørnstad for pointing out errors (and their corrections!) in the paper.

REFERENCES

AITKIN, M. and STASINOPOULOS, M. (1989). Likelihood analysis of a binomial sample size problem. In *Contributions to Probability and Statistics* (L. J. Gleser, M. D. Perlman, S. J. Press and A. Sampson, eds.) Springer, New York.

BARNARD, G. A., JENKINS, G. M. and WINSTEN, C. B. (1962). Likelihood inference and time series (with discussion). *J. Roy. Statist. Soc. Ser. A* **125** 321–372.

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

BARNDORFF-NIELSEN, O. (1988). *Parametric Statistical Models and Likelihood. Lecture Notes in Statist.* **50**. Springer, New York.

BARNDORFF-NIELSEN, O. (1991). Likelihood theory. In *Statistical Theory and Modelling: In Honour of Sir D.R. Cox*. Chapman and Hall, London.

BARTLETT, M. (1937). Properties of sufficiency and statistical tests. *Proc. Roy. Soc. London Ser. A* **160** 268–282.

BASU, D. (1975). Statistical information and likelihood (with discussion). *Sankhyā Ser. A* **37** 1–71.

BASU, D. (1977). On the elimination of nuisance parameters. *J. Amer. Statist. Assoc.* **72** 355–366.

BAYARRI, M. J., DEGROOT, M. H. and KADANE, J. B. (1988). What is the likelihood function? In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.) **2** 3–27. Springer, New York,

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, New York.

BERGER, J. O. and BERNARDO, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84** 200–207.

BERGER, J. O. and BERNARDO, J. M. (1992). Ordered group reference priors with applications to a multinomial problem. *Biometrika* **79** 25–37.

BERGER, J. O. and BERRY, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist* **76** 159–165.

BERGER, J. O., PHILIPPE, A. and ROBERT, C. (1998). Estimation of quadratic functions: noninformative priors for non-centrality parameters. *Statist. Sinica* **8** 359–376.

BERGER, J. O. and STRAWDERMAN, W. (1996). Choice of hierarchical priors: admissibility in estimation of normal means. *Ann. Statist.* **24** 931–951.

BERGER, J. O. and WOLPERT, R. L. (1988). *The Likelihood Principle: A Review, Generalizations, and Statistical Implications*, 2nd ed. IMS, Hayward, CA.

BERNARDO, J. M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 113–147.

BERNARDO, J. M. and SMITH, A. F. M. (1994). *Bayesian Theory*. Wiley, New York.

BJØRNSTAD, J. (1996). On the generalization of the likelihood function and the likelihood principle. *J. Amer. Statist. Assoc.* **91** 791–806.

BUTLER, R. W. (1988). A likely answer to "What is the likelihood function?" In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.) **2** 21–26. Springer, New York.

CARROLL, R. J. and LOMBARD, F. (1985). A note on N estimators for the Binomial distribution. *J. Amer. Statist. Assoc.* **80** 423–426.

CHANG, T. and EAVES, D. (1990). Reference priors for the orbit in a group model. *Ann. Statist.* **18** 1595–1614.

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276.

COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **49** 1–39.

CRUDDAS, A. M., REID, N. and COX, D. R. (1989). A time series illustration of approximate conditional likelihood. *Biometrika* **76** 231.

DATTA, G. S. and GHOSH, J. K. (1995a). Noninformative priors for maximal invariant parameter in group models. *Test* **4** 95–114.

DATTA, G. S. and GHOSH, J. K. (1995b). On priors providing frequentist validity for Bayesian inference. *Biometrika* **82** 37–45.

DATTA, G. S. and GHOSH, J. K. (1996). On the invariance of noninformative priors. *Ann. Statist.* **24** 141–159.

DAWID, A. P., STONE, M. and ZIDEK, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. Roy. Statist. Soc. Ser. B* **35** 180–233.

DE ALBA, E. and MENDOZA, M. (1996). A discrete model for Bayesian forecasting with stable seasonal patterns. In *Advances in Econometrics II* (R. Carter Hill, ed.) 267–281. JAI Press.

DRAPER, N. and GUTTMAN, I. (1971). Bayesian estimation of the binomial parameter. *Technometrics* **13** 667–673.

EATON, M. L. (1989). *Group Invariance Applications in Statistics*. IMS, Hayward, CA.

FISHER, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10** 507.

FISHER, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* **1** 3–32.

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

FRASER, D. A. S. and REID, N. (1989). Adjustments to profile likelihood. *Biometrika* **76** 477–488.

GHOSH, J. K., ed. (1988). *Statistical Information and Likelihood. A Collection of Critical Essays by D. Basu.* Springer, New York.

GHOSH, J. K. and MUKERJEE, R. (1992). Noninformative priors. In *Bayesian Statistics 4* (J. O. Berger, J. M. Bernardo, A. P. Dawid and A. F. M. Smith, eds.) 195–203. Oxford Univ. Press.

GLESER, L. and HWANG, J. T. (1987). The nonexistence of $100(1-\alpha)\%$ confidence sets of finite expected diameter in errors-in-variable and related models. *Ann. Statist.* **15** 1351–1362.

GOOD, I. J. (1983). *Good Thinking: The Foundations of Probability and Its Applications.* Univ. Minnesota Press.

HUI, S. and BERGER, J. O. (1983). Empirical Bayes estimation of rates in longitudinal studies. *J. Amer. Statist. Assoc.* **78** 753–760.

JEFFREYS, H. (1961). *Theory of Probability*. Oxford Univ. Press.

KAHN, W. D. (1987). A cautionary note for Bayesian estimation of the binomial parameter n. *Amer. Statist.* **41** 38–39.

KALBFLEISCH, J. D. and SPROTT, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *J. Roy. Statist. Soc. Ser. B* **32** 175–208.

KALBFLEISH, J. D. and SPROTT, D. A. (1974). Marginal and conditional likelihood. *Sankhyā Ser. A* **35** 311–328.

LAPLACE, P. S. (1812). *Theorie Analytique des Probabilities* Courcier, Paris.

LAVINE, M. and WASSERMAN, L. A. (1992). Can we estimate N? *Technical Report* 546, Dept. Statistics, Carnegie Mellon Univ.

LISEO, B. (1993). Elimination of nuisance parameters with reference priors. *Biometrika* **80** 295–304.

MCCULLAGH, P. and TIBSHIRANI, R. (1990). A simple method for the adjustment of profile likelihoods. *J. Roy. Statist. Soc. Ser. B* **52** 325–344.

MORENO, E. and GIRÓN, F. Y. (1995). Estimating with incomplete count data: a Bayesian Approach. Technical report, Univ. Granada, Spain.

NEYMAN, J. and SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32.

OLKIN, I., PETKAU, A. J. and ZIDEK, J. V. (1981). A comparison of n estimators for the binomial distribution. *J. Amer. Statist. Assoc.* **76** 637–642.

RAFTERY, A. E. (1988). Inference for the binomial N parameter: a hierarchical Bayes approach. *Biometrika* **75** 223–228.

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

REID, N. (1996). Likelihood and Bayesian approximation methods. In *Bayesian Statistics 5* (J. O. Berger, J. M. Bernardo, A. P. Dawid and A. F. M. Smith, eds.) 351–369. Oxford Univ. Press.

RISSANEN, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11** 416–431.

SAVAGE, L. J. (1976). On rereading R. A. Fisher. *Ann. Statist.* **4** 441–500.

SUN, D. (1994). Integrable expansions for posterior distributions for a two parameter exponential family. *Ann. Statist.* **22** 1808–1830.

SUN, D. and BERGER, J. O. (1998). Reference priors with partial information. *Biometrika* **85** 55–71.

SWEETING, T. (1995a). A framework for Bayesian and likelihood approximations in statistics. *Biometrika* **82** 1–24.

SWEETING, T. (1995b). A Bayesian approach to approximate conditional inference. *Biometrika* **82** 25–36.

SWEETING, T. (1996). Approximate Bayesian computation based on signed roots of log-density ratios. In *Bayesian Statistics 5* (J. O. Berger, J. M. Bernardo, A. P. Dawid and A. F. M. Smith, eds.) 427–444. Oxford Univ. Press.

YE, K. and BERGER, J. O. (1991). Non-informative priors for inference in exponential regression models. *Biometrika* **78** 645–656.

ZABELL, S. L. (1989). R.A. Fisher on the history of inverse probability. *Statist. Sci.* **4** 247–263.