Biometrika (2007), **94**, 3, *pp.* 529–542
© 2007 Biometrika Trust
Printed in Great Britain

doi:10.1093/biomet/asm040
Advance Access publication 5 August 2007

Integrated likelihood functions for non-Bayesian inference

BY THOMAS A. SEVERINI

Department of Statistics, Northwestern University, Evanston, Illinois 60208-4070, U.S.A.
severini@northwestern.edu

SUMMARY

Consider a model with parameter $\theta = (\psi, \lambda)$, where ψ is the parameter of interest, and let $L(\psi, \lambda)$ denote the likelihood function. One approach to likelihood inference for ψ is to use an integrated likelihood function, in which λ is eliminated from $L(\psi, \lambda)$ by integrating with respect to a density function $\pi(\lambda|\psi)$. The goal of this paper is to consider the problem of selecting $\pi(\lambda|\psi)$ so that the resulting integrated likelihood function is useful for non-Bayesian likelihood inference. The desirable properties of an integrated likelihood function are analyzed and these suggest that $\pi(\lambda|\psi)$ should be chosen by finding a nuisance parameter ϕ that is unrelated to ψ and then taking the prior density for ϕ to be independent of ψ. Such an unrelated parameter is constructed and the resulting integrated likelihood is shown to be closely related to the modified profile likelihood.

Some key words: Modified profile likelihood; Nuisance parameter; Orthogonal parameters; Reference prior.

1. INTRODUCTION

Consider a model with parameter $\theta \in \Theta$ and consider a parameter of interest $\psi = \varphi(\theta)$, taking values in a set Ψ. Let $L(\theta)$ denote the likelihood function corresponding to a particular set of data and consider likelihood inference for ψ. If $\varphi(\theta) = \theta$, then inference can be based directly on $L(\theta)$. However, when there is a nuisance parameter in the model, likelihood inference is often based on a pseudolikelihood function, a function of ψ and the data with properties similar to those of a likelihood function. For $\psi \in \Psi$, let $\Theta(\psi) = \{\theta \in \Theta : \varphi(\theta) = \psi\}$; corresponding to $\psi \in \Psi$ is the set of likelihoods $\mathcal{L}_\psi = \{L(\theta) : \theta \in \Theta(\psi)\}$. The goal in choosing a pseudolikelihood function is to summarize $\mathcal{L}_\psi$.

In simple cases, there a nuisance parameter λ such that the likelihood function for $\theta = (\psi, \lambda)$ may be written $L(\theta) = L_1(\psi)L_2(\lambda)$; in these cases, $L_1(\psi)$ can be used as a likelihood for ψ. However, in general, construction of a pseudolikelihood function depends on somewhat arbitrary considerations as to what constitutes an effective summary of $\mathcal{L}_\psi$. One commonly-used pseudolikelihood is the profile likelihood function, given by $L_p(\psi) = \sup_{\theta \in \Theta(\psi)} L(\theta)$; in this approach $\mathcal{L}_\psi$ is summarized by its maximum value.

Although use of $L_p(\psi)$ leads to methods with certain types of large-sample optimality, in small or moderate samples it is well known that $L_p(\psi)$ is not always an effective summary of $\mathcal{L}_\psi$; see for example Barndorff-Nielsen & Cox (1994, Chs. 3, 4), Berger et al. (1999), Kalbfleisch & Sprott (1970) and Severini (2000, Ch. 4) for further discussion of $L_p(\psi)$ and its properties. To deal with this fact, many alternatives to the profile likelihood have been proposed. One broad class of pseudolikelihood functions involves modification of $L_p(\psi)$ to

incorporate additional information; see Barndorff-Nielsen (1983, 1994), Barndorff-Nielsen & Cox (1994, Ch. 8), Cox & Reid (1987), Fraser (2003), Fraser & Reid (1989), Kalbfleisch & Sprott (1970, 1973), McCullagh & Tibshirani (1990) and Severini (2000, Ch. 9) for discussion of various approaches to modifying the profile likelihood function.

An alternative approach is to summarize $\mathcal{L}_\psi$ by its average value with respect to some weight function over $\Theta(\psi)$. Suppose that $\theta = (\psi, \lambda)$, where $\lambda \in \Lambda$, and let $\pi(\lambda|\psi)$ denote a nonnegative function on Λ. We will refer to $\pi(\lambda|\psi)$ as the conditional prior density for λ given ψ even though, for our purposes, it is not necessary that π be a genuine density function. Then the integrated likelihood function with respect to π is given by

$$\int_\Lambda L(\psi, \lambda)\pi(\lambda|\psi)d\lambda; \tag{1}$$

see for example Berger et al. (1999), Kalbfleisch & Sprott (1970), Liseo (1993) and Severini (2000, Ch. 8) for further discussion of integrated likelihoods.

Integrated likelihoods have the advantage that, unlike marginal and conditional likelihoods, they are always available and, unlike pseudolikelihoods based on $L_p(\psi)$, they are based on averaging rather than maximization. The primary drawback of the integrated likelihood approach is that the prior density must be chosen. The main focus of this paper is selection of $\pi(\lambda|\psi)$ so that the resulting integrated likelihood is useful for non-Bayesian likelihood inference and the study of the properties of that integrated likelihood.

2. INTEGRATED LIKELIHOOD FUNCTIONS

Before considering selection of $\pi(\lambda|\psi)$, we must deal with the issue of standardization. Assume that either

$$\int_\Lambda \pi(\lambda|\psi)d\lambda \tag{2}$$

has the same finite value for each ψ or, if (2) is infinite for each ψ, then a normalization procedure along the lines of that used by Bernardo (1979) and Berger & Bernardo (1989, 1992) in the development of the reference prior has been used; see Berger et al. (1999).

In order for an integrated likelihood function to be useful for non-Bayesian likelihood inference for ψ, there are a number of generally desirable properties that it should possess and these properties have implications for the selection of $\pi(\lambda|\psi)$. In this section, several such properties are considered and these suggest that $\pi(\lambda|\psi)$ should be chosen so that, if γ is a nuisance parameter that is 'unrelated' to ψ, then γ and ψ should be independent under $\pi(\lambda|\psi)$. Here we take 'unrelated' to mean that $\hat{\gamma}_\psi$, the maximum likelihood estimator of γ for fixed ψ, is approximately constant as a function of ψ; see for example Cox & Reid (1987).

To be more precise, γ will be said to be weakly unrelated to ψ if $\hat{\gamma}_\psi = \hat{\gamma} + O(n^{-1})$ for moderate deviations of ψ, that is, for ψ of the form $\psi = \hat{\psi} + O(n^{-1/2})$; here $\hat{\psi}$ denotes the maximum likelihood estimator of ψ. For instance, if γ and ψ are orthogonal parameters, then γ is weakly unrelated to ψ; see for example Cox & Reid (1987), Barndorff-Nielsen & Cox (1994, Ch. 2) and Pace & Salvan (1997, Ch. 4). If, in addition, $\hat{\gamma}_\psi = \hat{\gamma} + O(n^{-1/2})$ for large deviations of ψ, that is, for fixed values of ψ so that $\psi - \hat{\psi} = O(1)$, we will say that γ and ψ are strongly unrelated. Thus, γ and ψ are strongly unrelated if

$\hat{\gamma}_\psi = \hat{\gamma} + O(n^{-1/2})O(|\psi - \hat{\psi}|)$. Note that an orthogonal nuisance parameter is not, in general, strongly unrelated to ψ.

We now consider several desirable properties of an integrated likelihood. Suppose that there exists a nuisance parameter γ such that $L(\theta) = L_1(\psi)L_2(\gamma)$; then $L_1(\psi)$ can be used as a likelihood for ψ (Barnard et al., 1962, § 8). It is easy to see that the integrated likelihood will correspond to $L_1(\psi)$ provided that $\pi(\lambda|\psi)$ is such that ψ and γ are independent. Since $\hat{\gamma}_\psi$ does not depend on ψ, this property suggests that, for non-Bayesian inference about ψ, $\pi(\lambda|\psi)$ should be chosen so that unrelated parameters are independent.

If an integrated likelihood is to be used for non-Bayesian inference, then it should possess the frequentist properties of a genuine likelihood function, at least approximately. Important properties of this type are the first two Bartlett identities. Let $M(\psi)$ denote a genuine likelihood function for ψ and let $m(\psi) = \log M(\psi)$. The first Bartlett identity states that $E\{m'(\psi); \theta\} = 0$, a property known as score unbiasedness. Here, and throughout the paper, differentiation of a loglikelihood or pseudo-loglikelihood function depending only on ψ will be denoted by $'$, $''$ and so on; derivatives of the loglikelihood function $\ell(\psi, \lambda)$ will be denoted by subscripts so that, for example, $\ell_\lambda(\psi, \lambda) = \partial\ell(\psi, \lambda)/\partial\lambda$. The second Bartlett identity states that $E\{m''(\psi) + m'(\psi)m'(\psi)^{\mathrm{T}}; \theta\} = 0$, a property known as information unbiasedness.

If $\bar{L}(\psi)$ is an integrated likelihood function and $\bar{\ell}(\psi) = \log \bar{L}(\psi)$, then, in general, $E\{\bar{\ell}'(\psi); \theta\}$ and $E\{\bar{\ell}'(\psi) + \bar{\ell}'(\psi)\bar{\ell}'(\psi)^{\mathrm{T}}; \theta\}$ are both $O(1)$ as $n \to \infty$ (Severini, 1998b). However, if the model is parameterized by a nuisance parameter γ that is weakly unrelated to ψ and $\pi(\gamma|\psi)$ does not depend on ψ, then $E\{\bar{\ell}'(\psi); \psi, \lambda\} = O(n^{-1})$; see Sweeting (1987) and Severini (1998b). If γ is strongly unrelated to ψ and $\pi(\gamma|\psi)$ does not depend on ψ, then $E\{\bar{\ell}''(\psi) + \bar{\ell}'(\psi)\bar{\ell}'(\psi)^{\mathrm{T}}; \theta\}$ is also $O(n^{-1})$ (Severini, 1998b). Thus, this analysis suggests that $\pi(\lambda|\psi)$ should be chosen so that, if a nuisance parameter γ is strongly unrelated to ψ, then ψ and γ are independent under $\pi(\lambda|\psi)$.

Since the choice of prior density is, to some extent, arbitrary, the integrated likelihood should be insensitive to the choice of prior. Stated another way, we do not want the integrated likelihood function to depend heavily on incidental features of the prior. For instance, if $L(\theta) = L_1(\psi)L_2(\gamma)$, then any prior density $\pi(\lambda|\psi)$ under which ψ and γ are independent yields the same integrated likelihood. More generally, using a Laplace approximation, we have

$$\bar{L}(\psi) = \int_\Lambda L(\psi, \lambda)\pi(\lambda|\psi)d\lambda = cL_p(\psi)|-\ell_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{-1/2}\pi(\hat{\lambda}_\psi|\psi)\{1 + O(n^{-1})\},$$

where c denotes a constant not depending on ψ and $\ell_{\lambda\lambda}(\psi, \lambda) = \partial^2\ell(\psi, \lambda)/\partial\lambda\partial\lambda^{\mathrm{T}}$. Hence, to the order considered, $\bar{L}(\psi)$ depends on $\pi(\lambda|\psi)$ only through $\pi(\hat{\lambda}_\psi|\psi)$. Thus, if the model is parameterized in terms of a nuisance parameter γ that is weakly unrelated to ψ and $\pi(\gamma|\psi)$ does not depend on ψ, then, for ψ in the moderate deviation range, $\bar{L}(\psi)$ does not depend on the form of $\pi(\gamma)$, if terms of order n^{-1} are ignored. If γ is strongly unrelated to ψ, then, in addition, $\bar{L}(\psi)$ does not depend on the form of $\pi(\gamma)$ for ψ in the large deviation range, if terms of order $n^{-1/2}$ are ignored. Again, this property suggests that the prior should be chosen so that parameters that are strongly unrelated are independent.

Finally, the integrated likelihood function should be invariant with respect to interest-respecting reparameterizations of the model; see for example Barndorff-Nielsen & Cox (1994, Ch. 8) and Pace & Salvan (1997, Ch. 4).

The analysis in this section suggests that, in order to construct an integrated likelihood function that is useful for non-Bayesian inference, we can construct a nuisance parameter

ϕ that is strongly unrelated to ψ and then choose a prior density for ϕ that does not depend on ψ. Let $\tilde{L}(\psi, \phi)$ denote the likelihood function in terms of (ψ, ϕ). Then the integrated likelihood for ψ with respect to a density $\pi(\phi)$ for ϕ is given by

$$\bar{L}(\psi) = \int \tilde{L}(\psi, \phi)\pi(\phi)d\phi. \tag{3}$$

3. CONSTRUCTION OF A NUISANCE PARAMETER THAT IS UNRELATED TO THE PARAMETER OF INTEREST

Define a data-dependent parameter $\phi \equiv \phi(\psi, \lambda; \hat{\psi})$ by the solution to the implicit equation

$$E\{\ell_\lambda(\psi, \lambda); \hat{\psi}, \phi\} \equiv E\{\ell_\lambda(\psi, \lambda); \psi_0, \lambda_0\} \Big|_{(\psi_0, \lambda_0) = (\hat{\psi}, \phi)} = 0; \tag{4}$$

that is, fixing the value of $(\psi, \lambda, \hat{\psi})$ and solving (4) for ϕ yields $\phi(\psi, \lambda; \hat{\psi})$. In Appendix 1 it is shown that $\hat{\phi} = \hat{\lambda}$ and that ϕ is strongly unrelated to ψ. The function ϕ will be called the zero-score-expectation parameter.

An important feature of ϕ is that it depends on the data. While such a parameter clearly would not be useful in parameterizing a statistical model in the usual sense, using it to parameterize a likelihood function does not cause difficulties since, in the likelihood function, the data are considered fixed. Of course, in determining the frequentist properties of the resulting integrated likelihood it is important to keep it in mind that ϕ depends on the data and, for Bayesian inference, such a data-dependent parameter raises other important issues that do not arise in the non-Bayesian setting considered here.

The parameter ϕ can be derived using the following argument. Our goal is to find a function ϕ of (ψ, λ) such that $\hat{\phi}_\psi = \hat{\phi} + O(n^{-1/2})O(|\psi - \hat{\psi}|)$. Suppose that there exists such a parameter $\phi = g(\psi, \lambda)$; we may also write $\lambda = h(\psi, \phi)$. Then $\hat{\lambda}_\psi = \lambda(\psi, \hat{\phi}_\psi) = \lambda(\psi, \hat{\phi}) + O(n^{-1/2})$.

Thus, there exists a statistic $\hat{\phi}$, with the same dimension as λ, such that, for any ψ, $\hat{\lambda}_\psi$ depends on the data only through $\hat{\phi}$, to the order given. Although in some models, such as certain exponential family models, this is possible, in general such a statistic does not exist. Hence, we allow ϕ to be a function of $\hat{\psi}$ as well as of (ψ, λ), writing $\phi = g(\psi, \lambda; \hat{\psi})$ and $\lambda = h(\psi, \phi; \hat{\psi})$. Since $\hat{\lambda}_\psi = h(\psi, \hat{\phi}_\psi; \hat{\psi})$, if $\hat{\phi}_\psi = \hat{\phi} + O(n^{-1/2})O(|\psi - \hat{\psi}|)$, then we must have

$$\hat{\lambda}_\psi = h(\psi, \hat{\phi}; \hat{\psi}) + O(n^{-1/2})O(|\psi - \hat{\psi}|); \tag{5}$$

thus, we want to find a function h such that (5) holds.

First suppose that $(\hat{\psi}, \hat{\lambda})$ is sufficient. Then, if ϕ can be chosen so that $\hat{\phi} = \hat{\lambda}$, then the function h can be taken to be the function that expresses $\hat{\lambda}_\psi$ in terms of the sufficient statistic $(\hat{\psi}, \hat{\lambda})$ and ψ. Note that, if we write $\ell(\psi, \lambda)$ as $\ell(\psi, \lambda; \hat{\psi}, \hat{\lambda})$, $\hat{\lambda}_\psi$ is the function of $(\psi, \hat{\lambda}, \hat{\psi})$ given by the likelihood equation $\ell_\lambda(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\lambda}) = 0$. Thus, if $g(\psi, \lambda; \hat{\psi})$ is taken to be the solution in ϕ, and $h(\psi, \phi; \hat{\psi})$ is taken to be the solution in λ to the equation $\ell_\lambda(\psi, \lambda; \hat{\psi}, \phi) = 0$, then $\hat{\phi} = \hat{\lambda}$ and $\hat{\lambda}_\psi = h(\psi, \hat{\phi}; \hat{\psi})$, so that (5) holds. Hence, with this definition of ϕ, $\hat{\phi}_\psi = \hat{\phi}$ for all ψ.

Now suppose that $(\hat{\psi}, \hat{\lambda})$ is not necessarily sufficient. If $L(\psi, \lambda)$ depends on the data through $(\hat{\psi}, \hat{\lambda}, a)$, where a is an ancillary, and $\phi \equiv \phi(\psi, \lambda; \hat{\psi}, a)$ is defined to be the solution to $\ell_\lambda(\psi, \lambda; \hat{\psi}, \phi, a) = 0$, then we still have $\hat{\phi}_\psi = \hat{\phi}$ for all ψ. However, this is not

suitable as a general approach, because of the well known difficulties in constructing an ancillary statistic a such that $(\psi, \hat{\lambda}, a)$ is sufficient. Thus, in the general case, we replace $\ell(\psi, \lambda; \hat{\psi}, \hat{\lambda}, a)$ by $E\{\ell(\psi, \lambda); \hat{\psi}, \hat{\lambda}\}$, which leads to (4). In Appendix 1, it is shown that ϕ is strongly unrelated to ψ.

For those models for which the operation of expectation of $\ell(\psi, \lambda)$ with respect to the distribution with parameter $(\hat{\psi}, \hat{\lambda})$ yields $\ell(\psi, \lambda; \hat{\psi}, \hat{\lambda})$, the stronger property that $\hat{\phi}_\psi = \hat{\phi}$ for all ψ holds. For instance, this is the case for a full-rank exponential family model with loglikelihood of the form $\ell(\psi, \lambda) = c(\psi, \lambda)^{\mathrm{T}} x - d(\psi, \lambda)$; see Appendix 1.

Example 1: *Exponent of a Weibull distribution*. Let $Y_1, \ldots, Y_n$ denote independent, identically distributed random variables each distributed according to the distribution with density function $\psi\lambda(\lambda y)^{\psi-1} \exp\{-(\lambda y)^\psi\}$, $y > 0$, where $\psi > 0$ and $\lambda > 0$. Then

$$E\{\ell_\lambda(\psi, \lambda); \psi_0, \lambda_0\} = \frac{n\psi}{\lambda}\Big\{1 - \frac{\lambda^\psi}{\lambda_0^\psi}\Gamma(\psi/\psi_0 + 1)\Big\}$$

so that equation (4) used to define ϕ is given by

$$\frac{n\psi}{\lambda}\Big\{1 - \frac{\lambda^\psi}{\phi^\psi}\Gamma(\psi/\hat{\psi} + 1)\Big\} = 0.$$

It follows that the zero score-expectation parameter is given by $\phi = \lambda\, \Gamma(\psi/\hat{\psi} + 1)^{1/\psi}$.

Orthogonal nuisance parameters are of the form $g\{\log\lambda + (c-1)/\lambda\}$ where g is an arbitrary smooth function and c denotes Euler's constant. Thus, $\gamma = \lambda \exp\{(c-1)/\lambda\}$ is an orthogonal nuisance parameter.

Although both $\hat{\phi}_\psi$ and $\hat{\gamma}_\psi$ depend on ψ, $\hat{\gamma}_\psi$ depends on ψ to a greater extent than does $\hat{\phi}_\psi$. For instance, consider the data reported by Martz & Waller (1982, Example 4.13) on the failure times of two electrical circuits. Figure 1 gives plots of $(\hat{\phi}_\psi - \hat{\phi})/\hat{\sigma}_\phi$ and $(\hat{\gamma}_\psi - \hat{\gamma})/\hat{\sigma}_\gamma$, where $\hat{\sigma}_\phi$ and $\hat{\sigma}_\gamma$ denote the estimated standard errors of $\hat{\phi}$ and $\hat{\gamma}$, respectively, based on the 12 failure times of circuit 1, for a range of ψ values. This figure supports the claim that ϕ is less related to ψ than is γ.

4. Relationship between integrated likelihood and the modified profile likelihood

As noted in the introduction, there have been many proposals for modifying the profile likelihood in order to make it more useful for likelihood inference. Perhaps the most satisfactory of these is the modified profile likelihood, L_M, proposed by Barndorff-Nielsen (1983). The modified profile likelihood has a number of important properties: it arises as an approximation to a marginal or conditional likelihood when either is available, it satisfies the Bartlett identities to a high degree of approximation, and it is invariant under interest-respecting parameterizations (Barndorff-Nielsen & Cox, 1994, Ch. 8). In this section the relationship between L_M and the proposed class of integrated likelihood functions is considered.

Calculation of $L_M(\psi)$ is restricted to those models in which the sufficient statistic may be written $(\hat{\psi}, \hat{\lambda}, a)$, where a is ancillary. However, there are approximations to L_M that are available more generally. Let $\bar{L}_M(\psi)$ denote the approximation to $L_M(\psi)$ given by

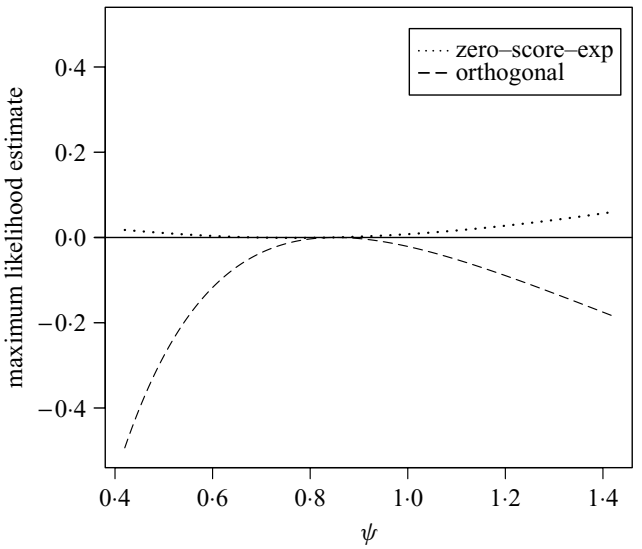

Fig. 1. Maximum likelihood estimates in Example 1.

Barndorff-Nielsen (1995) and Severini (1998a); see also Skovgaard (1996). Then

$$\bar{L}_M(\psi) = \frac{L_p(\psi)|\hat{j}_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{1/2}}{|I(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\lambda})|}, \tag{6}$$

where $\hat{j}_{\lambda\lambda}(\psi, \hat{\lambda}_\psi) = -\ell_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)$, and $I(\psi, \lambda; \psi_0, \lambda_0) = E\{\ell_\lambda(\psi, \lambda)\ell_\lambda(\psi_0, \lambda_0)^{\mathrm{T}}; \psi_0, \lambda_0\}$.

When considering an approximation to a pseudolikelihood function, we are generally interested in values of ψ near $\hat{\psi}$, that is, the moderate-deviation range of ψ. However, when we consider the properties of quantities derived from the pseudolikelihood, such as the cumulants of loglikelihood derivatives, the large-deviation properties often play a role. For instance, if $R_n(\psi) = O(n^{-1/2})$ for each ψ, then $\{R_n(\psi_1) - R_n(\psi)\}/(\psi_1 - \psi)$ is $O(n^{-1/2})$ for each ψ_1 and ψ. It follows that $dR_n(\psi)/d\psi = O(n^{-1/2})$; this result does not hold if $R_n(\psi) = O(n^{-1/2})$ only for ψ in the moderate-deviation range.

The function $\bar{L}_M(\psi)$ approximates $L_M(\psi)$ with error $O(n^{-1})$ for ψ in the moderate-deviation range and with error $O(n^{-1/2})$ for ψ in the large-deviation range (Severini, 1998a). Thus, we can write $\bar{L}_M(\psi) = cL_M(\psi)\{1 + O(n^{-1/2})O(|\psi - \hat{\psi}|)\}$, where c is a generic constant.

In Appendix 2 it is shown that $\bar{L}(\psi) = c\bar{L}_M(\psi)\{1 + O(n^{-1/2})O(|\psi - \hat{\psi}|)\}$. When the model is a full-rank exponential family model, then $\bar{L}(\psi) = c\bar{L}_M(\psi)\{1 + O(n^{-1})O(|\psi - \hat{\psi}|)\}$. If we use the relationship between $\bar{L}_M(\psi)$ and $L_M(\psi)$, it follows that $\bar{L}(\psi) = cL_M(\psi)\{1 + O(n^{-1/2})O(|\psi - \hat{\psi}|)\}$.

In the next section we consider the extent to which $\bar{L}(\psi)$ satisfies the properties discussed in § 2.

5. PROPERTIES OF THE INTEGRATED LIKELIHOOD

5·1. *Models in which the likelihood factorizes*

Suppose that the likelihood is of the form $L_1(\psi)L_2(\gamma)$ for some nuisance parameter γ and some functions L_1 and L_2. Then ϕ depends on ψ and λ only through $\gamma(\psi, \lambda)$ and,

hence, $\bar{L}(\psi)$ is proportional to $L_1(\psi)$ for any choice of $\pi(\phi)$ such that the integral in (3) is finite.

5·2. *Frequentist properties: score and information bias*

Let $\bar{\ell}(\psi)$ and $\bar{\ell}_M(\psi)$ denote $\log \bar{L}(\psi)$ and $\log \bar{L}_M(\psi)$, respectively. Then $E\{\bar{\ell}'_M(\psi); \psi, \lambda\}$ and $E\{\bar{\ell}''_M(\psi); \psi, \lambda\} + E\{\bar{\ell}'_M(\psi)\bar{\ell}'_M(\psi)^{\mathrm{T}}; \psi, \lambda\}$ are both $O(n^{-1})$, and the relationship between $\bar{\ell}(\psi)$ and $\bar{\ell}_M(\psi)$ implies that $E\{\bar{\ell}'(\psi); \psi, \lambda\}$ and $E\{\bar{\ell}''(\psi); \psi, \lambda\} + E\{\bar{\ell}'(\psi)\bar{\ell}'(\psi)^{\mathrm{T}}; \psi, \lambda\}$ are both $O(n^{-1})$; see Ferguson et al. (1991), DiCiccio et al. (1996) and Severini (1998b). It follows that $\bar{L}(\psi)$ is score-unbiased and information-unbiased to order $O(n^{-1})$.

Thus, since the proposed integrated likelihood function can be approximated by $\bar{L}_M$, it has the same desirable frequentist properties as $\bar{L}_M$. Conversely, $\bar{L}_M$ provides a higher-order approximation to any integrated likelihood function that has good frequentist properties. This conclusion follows from the following argument. Consider an integrated likelihood based on an arbitrary prior for λ; denote this integrated likelihood by $G(\psi)$. By changing the variable of integration to ϕ, we may write

$$Q(\psi) = \int \tilde{L}(\psi, \phi)\pi(\phi|\psi; \hat{\psi})d\phi.$$

Here $\pi(\phi|\psi; \hat{\psi})$ denotes the corresponding prior for ϕ; note that, since ϕ is a function of $(\psi, \lambda, \hat{\psi})$, this prior generally depends on $\hat{\psi}$.

Using a Laplace approximation as in Appendix 2, we obtain

$$Q(\psi) = c\bar{L}_M(\psi)\pi(\hat{\phi}_\psi|\psi; \hat{\psi})\{1 + D_n(\psi)\}, \tag{7}$$

where $D_n(\psi) = O(n^{-1})$ for any fixed ψ and c does not depend on ψ. Let $w(\phi, \psi; \hat{\psi}) = \log \pi(\phi|\psi; \hat{\psi})$ and let subscripts on w denote differentiation. Note that, by the properties of $\hat{\phi}_\psi$ described in Appendix 1,

$$\frac{d}{d\psi}w(\hat{\phi}_\psi, \psi; \hat{\psi}) = w_\phi(\hat{\phi}_\psi, \psi; \hat{\psi})\hat{\phi}'_\psi + w_\psi(\hat{\phi}_\psi, \psi; \hat{\psi}) = w_\psi(\lambda, \psi; \hat{\psi}) + O_p(n^{-1/2}). \tag{8}$$

Suppose that $E\{q'(\psi); \psi, \lambda\} = O(n^{-1})$, where $q = \log Q$. Since $E\{\bar{\ell}'_M(\psi); \psi, \lambda\} = O(n^{-1})$ (Severini, 1998b) and, from (8),

$$E\{\frac{d}{d\psi}w(\hat{\phi}, \psi; \hat{\psi}); \lambda, \psi\} = w_\psi(\lambda, \psi; \psi) + O(n^{-1/2}),$$

it follows that $w_\psi(\lambda, \psi; \psi) = 0$ for all λ and ψ. Hence, for $\psi = \hat{\psi} + O(n^{-1/2})$, $w(\hat{\phi}_\psi, \psi; \hat{\psi}) = w(\hat{\lambda}, \hat{\psi}; \hat{\psi}) + O_p(1/n)$ so that $\pi(\hat{\phi}_\psi|\psi; \hat{\psi}) = \pi(\hat{\lambda}|\hat{\psi}; \hat{\psi})\{1 + O(n^{-1})\}$.

It now follows from (7) that $Q(\psi)$ can be approximated by $\bar{L}_M(\psi)$, or by $L_M(\psi)$, with error $O(n^{-1})$ for ψ in the moderate-deviation range; that is, any integrated likelihood function $Q(\psi)$, that is approximately score-unbiased, is approximately equal to the modified profile likelihood. A similar result, expressed differently, is given by Severini (1998b).

Similarly, the fact that $Q(\psi)$ is information-unbiased to $O(n^{-1})$ suggests that $Q(\psi)$ can be approximated by $\bar{L}_M(\psi)$ with error $O(n^{-1/2})$ for fixed ψ. However, a proof of this result is not available since approximate information-unbiasedness only puts conditions on the second derivatives of w and these are not sufficient to establish large-deviation properties.

5·3. *Dependence on the prior density used*

Since, for any choice of prior $\pi(\phi)$, the proposed integrated likelihood function can be approximated by $\bar{L}_M(\psi)$, clearly $\bar{L}(\psi)$ does not heavily depend on the specific prior used. More precisely, let π_1 and π_2 denote priors for ϕ and let $\bar{L}^{(j)}(\psi)$, $j = 1, 2$, denote the corresponding integrated likelihoods. Then, if we neglect terms that do not depend on ψ, $\bar{L}^{(1)}(\psi)/\bar{L}^{(2)}(\psi)$ is $1 + O(n^{-1/2})$, for fixed ψ, and is $1 + O(n^{-1})$, for $\psi = \hat{\psi} + O(n^{-1/2})$. For full-rank exponential family models, the ratio is $1 + O(n^{-1})$ for fixed ψ and is $1 + O(n^{-3/2})$ for $\psi = \hat{\psi} + O(n^{-1/2})$.

For comparison, the ratio of arbitrary integrated likelihoods based on prior densities $\pi_1(\lambda|\psi)$ and $\pi_2(\lambda|\psi)$ is $O(1)$ for fixed ψ and is $1 + O(n^{-1/2})$ for $\psi = \hat{\psi} + O(n^{-1/2})$. If ψ and λ are orthogonal parameters and the prior densities do not depend on ψ, then the ratio is $O(1)$ for fixed ψ and $1 + O(n^{-1})$ for $\psi = \hat{\psi} + O(n^{-1/2})$.

5·4. *Parameterization invariance*

There are two aspects to the invariance of $\bar{L}(\psi)$ under reparameterizations leaving ψ unchanged. One is the effect of reparameterization on the definition of the zero score-expectation parameter ϕ. Let $\eta \equiv \eta(\psi, \lambda)$ denote an alternative nuisance parameter for the model and let $\omega \equiv \omega(\psi, \eta, \hat{\psi})$ denote the zero score-expectation parameter based on (ψ, η). Then $\omega(\psi, \eta, \hat{\psi}) = \phi\{\psi, \lambda(\psi, \eta), \hat{\psi}\}$, where ϕ is based on the original parameterization.

The second aspect of invariance is the properties of the prior density for ϕ. If the method of selecting the prior for ϕ is parameterization-invariant, then $\bar{L}(\psi)$ is parameterization-invariant. However, since neither the prior for ϕ nor the prior for ω can depend on ψ, and the function relating ϕ and ω depends on ψ, commonly-used parameterization-invariant priors, such as those discussed by Kass & Wasserman (1996), cannot generally be applied in this setting. The problem of constructing such a prior density will not be considered here.

For a fixed type of prior, such as a uniform prior, $\bar{L}(\psi)$ is parameterization-invariant to order $O(n^{-1/2})$ for fixed ψ and to order $O(n^{-1})$ for $\psi = \hat{\psi} + O(n^{-1/2})$. For comparison, an integrated likelihood function based on a orthogonal nuisance parameter is parameterization-invariant to $O(1)$ for fixed ψ and to order $O(n^{-1})$ for $\psi = \hat{\psi} + O(n^{-1/2})$.

6. EXAMPLES

Example 2: *Ratio of normal means*. Let X and Y denote independent normal random variables such that X has mean μ_1 and variance $1/n$ and Y has mean μ_2 and variance $1/n$. Let $\psi = \mu_1/\mu_2$ and take $\lambda = \mu_2$ as the nuisance parameter; assume that $\lambda \neq 0$. The profile loglikelihood is given by

$$\ell_p(\psi) = -\frac{n}{2}\frac{(x - \psi y)^2}{1 + \psi^2}, \quad -\infty < \psi < \infty.$$

Note that $\lim_{|\psi|\to\infty} \ell_p(\psi) = -ny^2/2$ so that, for certain values of n and y, the usual likelihood ratio confidence intervals for ψ are of infinite length and, possibly, the entire real line.

Here $E\{\ell_\lambda(\psi, \lambda); \psi_0, \lambda_0\} = n(\psi\psi_0 + 1)\lambda_0 - n(\psi^2 + 1)\lambda$ so that the zero-score-expectation parameter is given by $\phi = (\psi^2 + 1)\lambda/(\psi\hat{\psi} + 1)$. If $\pi(\phi)$ is taken to be the density of the normal distribution with mean 0 and variance $1/\tau^2$, then the integrated

loglikelihood is

$$\bar{\ell}(\psi;\tau^2)=\frac{n}{2}\frac{(\psi x+y)^2}{\psi^2+1}\left\{1+\frac{\tau^2}{n}\frac{\psi^2+1}{(\psi\hat{\psi}+1)^2}\right\}^{-1}+\frac{1}{2}\log(\psi^2+1)-\log|\psi\hat{\psi}+1|$$
$$+\log\left\{1+\frac{\tau^2}{n}\frac{\psi^2+1}{(\psi\hat{\psi}+1)^2}\right\}.$$

For this model, Liseo (1993) recommends the use of L_R, the integrated likelihood based on the reference prior for the nuisance parameter; see Berger et al. (1999). Here

$$\ell_R(\psi)=-\frac{n}{2}\frac{(x-\psi y)^2}{1+\psi^2}-\frac{1}{2}\log(1+\psi^2),\quad -\infty<\psi<\infty$$

(Liseo, 1993), where $\ell_R=\log L_R$. Then $\ell_R(\psi)$ approaches 0 as $|\psi|\to\infty$, so that likelihood-based interval estimates for ψ based on ℓ_R will be of finite length. Liseo (1993) and Berger et al. (1999) view this as a very desirable property of ℓ_R. However, if y is close to 0, it seems reasonable that it is not possible to rule out very large values of $E(X)/E(Y)$ on the basis of x and y alone. Thus, ℓ_R, which does rule out such very large values, must be including information other than that provided by the data and, hence, we would expect this to be reflected in the frequentist properties of procedures based on ℓ_R. For instance,

$$E\{\ell_R'(\psi);\psi,\lambda\}=-\frac{\psi}{1+\psi^2},\quad E\{\ell_R'(\psi)^2+\ell_R''(\psi);\psi,\lambda\}=2\frac{\psi^2}{(1+\psi^2)^2},$$

so that ℓ_R is both score- and information-biased.

A small Monte Carlo study was conducted to consider the coverage probability of confidence intervals of the form $2\{\ell(\hat{\psi})-\ell(\psi)\}\leqslant\chi^2_\alpha$ where $\ell(\psi)$ is one of $\bar{\ell}(\psi;1/2)$, $\ell_R(\psi)$ and $\ell_p(\psi)$, $\hat{\psi}$ denotes the maximizer of $\ell(\psi)$, and χ^2_α denotes the $1-\alpha$ quantile of the chi-squared distribution with 1 degree of freedom. Table 1 gives the estimated coverage probability of such intervals for the case $\alpha=0.05$, $\mu_1=4$ and $\mu_2=1/5$, for several choices of n. These results show that at least some frequentist properties of the proposed integrated likelihoods are superior to those of the reference integrated likelihood.

The modified profile loglikelihood function is given by

$$\ell_M(\psi)=\frac{n}{2}\frac{(\psi x+y)^2}{\psi^2+1}+\frac{1}{2}\log(\psi^2+1)-\log|\psi\hat{\psi}+1|,$$

which corresponds to $\bar{\ell}(\psi;\tau^2)$ with $\tau^2=0$; here $\bar{L}_M=L_M$. Note that $\ell_M(\psi)$ is maximized by $\hat{\psi}_M=-1/\hat{\psi}$, which is clearly an inconsistent estimator of ψ. For any $\tau^2>0$, $\bar{\ell}(\psi;\tau^2)$ has a local maximum at $\hat{\psi}_M$ and there is positive probability that the maximizer of $\bar{\ell}(\psi;\tau^2)$ is equal to $\hat{\psi}_M$; however, this probability approaches 0 rapidly as n increases and the maximizer of $\bar{\ell}(\psi;\tau^2)$ is a consistent estimator of ψ for any $\tau^2>0$.

Example 3: *Gamma distributions with common shape parameter*. For each $j=1,\ldots,m$, let $Y_{j1},\ldots,Y_{jn_j}$ be independent, each with a gamma distribution with mean ψ/λ_j and variance ψ/λ_j^2. Then $\phi=(\phi_1,\ldots,\phi_m)$, where $\phi_j=\hat{\psi}\lambda_j/\psi$. If the prior for $(\phi_1,\ldots,\phi_m)$ is of the form $\pi_0(\phi_1)\cdots\pi_0(\phi_m)$, then

$$\bar{L}(\psi)=\frac{\psi^{n\psi}(t_1\cdots t_m)^\psi}{\hat{\psi}^{n\psi}\Gamma(\psi)^n}\prod_{j=1}^m\int_0^\infty\phi_j^{n_j\psi}\exp\left(-\frac{\psi}{\hat{\psi}}\phi_j s_j\right)\pi_0(\phi_j)d\phi_j,$$

Table 1. *Coverage probabilities in Example* 2

	Likelihood		
n	L_p	L_R	$\bar{L}$
1	0·951	0·490	0·962
2	0·950	0·637	0·954
5	0·949	0·763	0·951
10	0·948	0·828	0·949
25	0·951	0·888	0·951
100	0·949	0·940	0·949

where $n = \sum_{j=1}^{m} n_j, t_j = y_{j1} \cdots y_{jn_j}$, and $s_j = \sum_{k=1}^{n_j} y_{jk}$. For instance, if either $\pi_0(\phi_j) = 1/\phi_j$ or $\pi_0(\phi_j) = 1$, then $\bar{L}(\psi)$ is identical to the conditional likelihood given $(s_1, \ldots, s_m)$, $L_C(\psi)$. A comparison of $L_C(\psi)$ and $L_p(\psi)$ for a specific set of data is given in Severini (2000, Example 8.2).

The modified profile likelihood function for this model is given by

$$L_M(\psi) = L_C(\psi) \prod_{j=1}^{m} \frac{(n_j\psi)^{n_j\psi - 1/2} \exp(-n_j\psi)}{\Gamma(n_j\psi)}$$

and $\bar{L}_M = L_M$; note that $x^{x-1/2}\exp(-x)$ is Stirling's approximation to $\Gamma(x)$.

Example 4: *Normal distributions with common mean.* Let Y_{jk}, $k = 1, \ldots, n_j$, $j = 1, \ldots, m$, denote independent normal random variables such that Y_{jk} has mean μ and standard deviation σ_j. The parameter of interest is the common mean μ with $(\sigma_1, \ldots, \sigma_m)$ as the nuisance parameter. It is straightforward to show that $(\phi_1, \ldots, \phi_m)$ is given by $\phi_j^2 = \sigma_j^2 - (\hat{\mu} - \mu)^2$, $j = 1, \ldots, m$.

For the prior density $\pi(\phi_1, \ldots, \phi_m) = (\phi_1 \cdots \phi_m)^{-1/2}$ the integrated likelihood is

$$\bar{L}(\psi) = |\hat{\mu} - \mu|^{-(n-m)} \prod_{j=1}^{m} M\left(\frac{n_j - 1}{2}, \frac{n_j}{2}, -\frac{s_j(\mu)}{2(\hat{\mu} - \mu)^2}\right), \tag{9}$$

where $n = \sum_{j=1}^{m} n_j$, $s_j(\mu)^2 = \sum_{k=1}^{n_j} (y_{jk} - \mu)^2$ and M denotes the confluent hypergeometric function (Abramowitz & Stegun, 1965, Ch. 13). For this model, the profile likelihood is given by $L_p(\psi) = \prod_{j=1}^{m} s_j(\mu)^{-n_j}$ and $\bar{L}_M(\psi) = \prod_{j=1}^{m} s_j(\mu)^{-(n_j-2)}$.

For small sample sizes, inferences based on $\bar{L}(\psi)$ and $\bar{L}_M(\psi)$ may be quite different. For instance, if $n_j = 2$, then stratum j is not included in $\bar{L}_M(\psi)$, while its contribution to $\bar{L}(\psi)$ is

$$\exp\left\{-\frac{s_j(\mu)^2}{4(\hat{\mu} - \mu)^2}\right\} I_0\left\{\frac{s_j(\mu)^2}{4(\hat{\mu} - \mu)^2}\right\} / |\hat{\mu} - \mu|; \tag{10}$$

here I_0 denotes a modified Bessel function. It may be shown that, for small $|\hat{\mu} - \mu|$, (10) can be expanded as $\{2/s_j(\mu)\}\{1 + O(|\hat{\mu} - \mu|^2)\}$.

Consider Rayleigh's data on the mass of nitrogen, as reported by Jeffreys (1983, § 5·47). Here we consider the results from the three chemical methods used so that $m = 3$, $n_1 = 4$, $n_2 = 2$ and $n_3 = 2$. For convenience, the linear transformation $100(x - 2 \cdot 29)$ was used for each measurement. The observations from method j are assumed to be normally distributed with mean μ and standard deviation σ_j. Figure 2 contains plots of $\bar{L}(\psi)$, given by (9), $L_p(\psi)$ and $\bar{L}_M(\psi)$ for these data. Clearly, inferences based on $\bar{L}_M(\psi)$, which ignores the data from Methods 2 and 3, will be quite different from the inferences based on $\bar{L}(\psi)$.

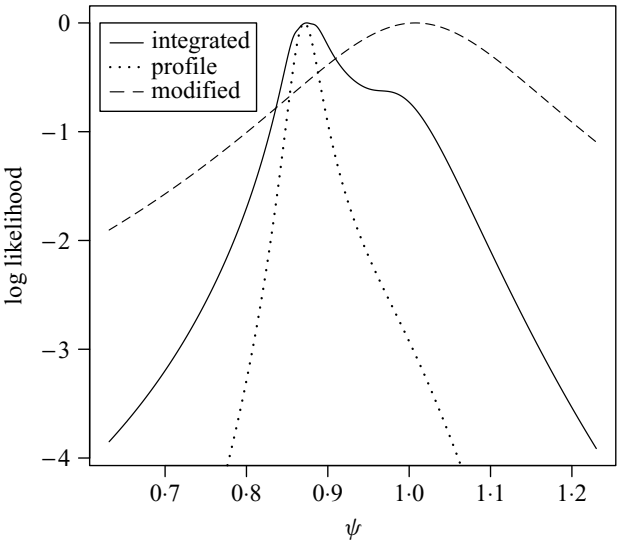

Fig. 2. Likelihood functions in Example 4.

7. Discussion

An important property of the proposed integrated likelihood functions is that they are closely related to the modified profile likelihood $L_M(\psi)$. There at least two ways to interpret this result. One is that the modified profile likelihood L_M, or its approximation $\bar{L}_M$, can be viewed as an approximation to a class of integrated likelihood functions that are useful for non-Bayesian inference so that the modified profile likelihood naturally arises from a non-Bayesian approach to integrated likelihood. Thus, this gives an interpretation of the modified profile likelihood that is valid when the condition used for its derivation, the existence of a marginal or conditional likelihood function, does not hold. Furthermore, calculation of L_M and $\bar{L}_M$ does not require selection of a prior density for the nuisance parameter, a major drawback of integrated likelihood functions.

A second interpretation of the result is that the proposed integrated likelihood functions can be used as approximations to the modified profile likelihood. In particular, it follows from the results in Appendix 2 that, if a uniform prior is used for ϕ, then the resulting integrated likelihood function $\bar{L}_U$ satisfies $\bar{L}_U(\psi) = \bar{L}_M(\psi)\{1 + O(n^{-1})O(|\psi - \hat{\psi}|)\}$ so that $\bar{L}_U$ can be expected to provide an accurate approximation to $\bar{L}_M$. Given that computation and approximation of integrals is a well-studied numerical problem (Evans & Swartz, 2000), computation of L_U may be preferable to computation of $\bar{L}_M$ in some models of interest. Furthermore, Examples 2 and 4 suggest that, in some cases, the proposed integrated likelihoods may have better properties than L_M and $\bar{L}_M$. This may be because $\bar{L}$ is based directly on averaging the likelihood function, rather than on higher-order distributional approximations.

Acknowledgement

I would like to thank the referees and associate editor for many useful comments and suggestions which greatly improved the paper. This work was supported by the U. S. National Science Foundation.

APPENDIX 1

Properties of the zero-score-expectation parameter

Since $\phi \equiv \phi(\psi, \lambda; \hat{\lambda})$ is the solution in ϕ of $E\{\ell_\lambda(\psi, \lambda); \hat{\psi}, \phi\} = 0$, it follows that $\hat{\phi}_\psi$ solves $E\{\ell_\lambda(\psi, \hat{\lambda}_\psi); \hat{\psi}, \hat{\phi}_\psi\} = 0$. In this and similar expressions, in computing the expectations any random variable taking the place of a parameter value in a loglikelihood or loglikelihood derivative will be assumed to be fixed; for example,

$$E\{\ell_\lambda(\psi, \hat{\lambda}_\psi); \hat{\psi}, \hat{\phi}_\psi\} \equiv E\{\ell_\lambda(\psi, \lambda); \hat{\psi}, \hat{\phi}_\psi\} \Big|_{\lambda=\hat{\lambda}_\psi}.$$

The proof that $\hat{\phi}_\psi = \hat{\lambda} + O(n^{-1/2})O(|\psi - \hat{\psi}|)$ uses the following idea. Let $\bar{\lambda}_\psi$ satisfy $E\{\ell_\lambda(\psi, \bar{\lambda}_\psi); \hat{\psi}, \hat{\lambda}\} = 0$ and let $\bar{\phi}_\psi \equiv \phi(\psi, \bar{\lambda}_\psi, \hat{\psi})$. Note that $\bar{\phi}_\psi$ solves $E\{\ell_\lambda(\psi, \bar{\lambda}_\psi); \hat{\psi}, \bar{\phi}_\psi\} = 0$; hence, it follows from the definition of $\bar{\lambda}_\psi$ that $\bar{\phi}_\psi = \hat{\lambda}$ for all ψ. The relationship between $\bar{\phi}_\psi$ and $\hat{\phi}_\psi$ can now be used to show that $\hat{\phi}_\psi = \hat{\lambda} + O(n^{-1/2})O(|\psi - \hat{\psi}|)$.

Let E_0 denote expectation with respect to the true distribution. Then, for any ψ, λ,

$$E\left\{\frac{1}{n}\ell_\lambda(\psi, \lambda); \hat{\psi}, \hat{\lambda}\right\} = E_0\left\{\frac{1}{n}\ell_\lambda(\psi, \lambda)\right\} + O_p\left(\frac{1}{\sqrt{n}}\right) = \frac{1}{n}\ell_\lambda(\psi, \lambda) + O_p\left(\frac{1}{\sqrt{n}}\right);$$

note that, since (ψ, λ) is not the true parameter value, $\ell_\lambda(\psi, \lambda) = O_p(n)$. Since $\hat{\lambda}_\psi$ satisfies $\ell_\lambda(\psi, \hat{\lambda}_\psi) = 0$, it follows that, for any value of ψ, $\bar{\lambda}_\psi = \hat{\lambda}_\psi + O_p(\frac{1}{\sqrt{n}})$; a similar argument can be used to show that $d\bar{\lambda}_\psi/d\psi = d\hat{\lambda}_\psi/d\psi + O_p(n^{-1/2})$. Thus, for any value of ψ, $\bar{\phi}_\psi = \hat{\phi}_\psi + O_p(n^{-1/2})$ and $d\hat{\phi}_\psi/d\psi = d\bar{\phi}_\psi/d\psi + O_p(n^{-1/2})$. The desired result now follows from the properties of $\bar{\phi}_\psi$.

Now consider a full-rank exponential family model with loglikelihood of the form $\ell(\psi, \lambda) = c(\psi, \lambda)^{\mathrm{T}} x - d(\psi, \lambda)$. Since $E(X; \theta) = \{c_\theta(\theta)\}^{-1} d_\theta(\theta)$, where subscripts denote differentiation,

$$E\{\ell_\lambda(\psi, \lambda); \psi_0, \lambda_0\} = \{c_\theta(\theta_0)^{-1} c_\lambda(\theta)\}^{\mathrm{T}} d_\theta(\theta_0) - d_\lambda(\theta), \quad \theta = (\psi, \lambda).$$

Since $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$ satisfies $c_\lambda(\hat{\theta}_\psi)^{\mathrm{T}} x - d_\lambda(\hat{\theta}_\psi) = 0$ and $\hat{\theta}$ satisfies $c_\theta(\hat{\theta})^{\mathrm{T}} x - d_\theta(\hat{\theta}) = 0$,

$$E\{\ell_\lambda(\psi, \hat{\lambda}_\psi); \hat{\psi}; \hat{\lambda}\} = \{c_\theta(\hat{\theta})^{-1} c_\lambda(\hat{\theta}_\psi)\}^{\mathrm{T}} d_\theta(\hat{\theta}) - d_\lambda(\hat{\theta}_\psi) = c_\lambda(\hat{\theta}_\psi)^{\mathrm{T}} x - d_\lambda(\hat{\theta}_\psi) = 0.$$

It follows that $E\{\ell_\lambda(\psi, \hat{\lambda}_\psi); \hat{\psi}, \hat{\lambda}\} = 0$ and hence that $\hat{\phi}_\psi = \hat{\phi}$ for all ψ.

APPENDIX 2

An expansion for the integrated likelihood function

Using a Laplace approximation (Evans & Swartz, 2000, § 4), we obtain

$$\bar{L}(\psi) = \int \tilde{L}(\psi, \phi)\pi(\phi)d\phi \; = c_0 \, \tilde{L}(\psi, \hat{\phi}_\psi)| - \tilde{\ell}_{\phi\phi}(\psi, \hat{\phi}_\psi)|^{-1/2} \pi(\hat{\phi}_\psi)\{1 + D_n(\psi)\} \qquad \text{(A1)}$$

for any given prior $\pi(\cdot)$; here $D_n(\psi) = O(n^{-1})$ for any fixed ψ and c_0 does not depend on ψ. Note that this approximation can be justified, since ϕ depends on $\hat{\psi}$, by changing the variable of integration to λ, which does not depend on the data, and then using a Laplace approximation on the resulting integral.

Since $\hat{\phi}_\psi = \hat{\lambda} + O(n^{-1/2})O(|\psi - \hat{\psi}|)$, $\pi(\hat{\phi}_\psi)/\pi(\hat{\phi}) = 1 + O(n^{-1/2})O(|\psi - \hat{\psi}|)$ and hence, if we let $c = c_0/\pi(\hat{\phi})$, it follows that

$$\bar{L}(\psi) = c \, \tilde{L}(\psi, \hat{\phi}_\psi)| - \tilde{\ell}_{\phi\phi}(\psi, \hat{\phi}_\psi)|^{-1/2}\{1 + O(n^{-1/2})O(|\psi - \hat{\psi}|)\}.$$

Since $\tilde{L}(\psi, \hat{\phi}_\psi) = L_p(\psi)$, these results show that $\bar{L}(\psi)$ can be approximated by $L_A(\psi) = \tilde{L}(\psi, \hat{\phi}_\psi)| - \tilde{\ell}_{\phi\phi}(\psi, \hat{\phi}_\psi)|^{-1/2}$. Note that $L_A(\psi)$ is the Cox-Reid adjusted profile likelihood (Cox & Reid, 1987) calculated using ϕ as the nuisance parameter; the fact that, in general, the

Cox-Reid adjusted profile likelihood is approximately an integrated likelihood was discussed by Sweeting (1987).

Thus, $\bar{L}(\psi) = c\bar{L}_M(\psi)\{1 + O(n^{-1/2})O(|\psi - \hat{\psi}|)\}$ provided that $L_A(\psi)$ and $L_M(\psi)$ agree to the required order. Clearly, a sufficient condition for this is that

$$|-\tilde{\ell}_{\phi\phi}(\psi, \hat{\phi}_\psi)|^{-1/2} = c_1 \frac{|\hat{j}_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{\frac{1}{2}}}{|I(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\lambda})|} \tag{A2}$$

to the required order, where c_1 depends only on the data.

The validity of A2 is a consequence of the following results:

$$\tilde{\ell}_{\phi\phi}(\psi, \phi) = \Big(\frac{\partial\lambda}{\partial\phi}(\psi, \phi)\Big)^{\mathrm{T}} \ell_{\lambda\lambda}\{\psi, \lambda(\psi, \phi)\}\Big(\frac{\partial\lambda}{\partial\phi}(\psi, \phi)\Big), \tag{A3}$$

$$\frac{\partial\lambda}{\partial\phi}(\psi, \phi) = E[-\ell_{\lambda\lambda}\{\psi, \lambda(\psi, \phi)\}; \hat{\psi}, \phi]^{-1} I\{\psi, \lambda(\psi, \phi); \hat{\psi}, \phi\}, \tag{A4}$$

$$\frac{|I(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\phi}_\psi)|}{|I(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\lambda})|} = 1 + O(n^{-1/2})O(|\psi - \hat{\psi}|), \tag{A5}$$

$$\frac{|E\{-\ell_{\lambda\lambda}(\psi, \hat{\lambda}_\psi); \hat{\psi}, \hat{\phi}_\psi\}|}{|-\ell_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|} = c_1\{1 + O(n^{-1/2})O(|\psi - \hat{\psi}|)\}. \tag{A6}$$

Equation A3 follows from the chain rule for differentiation and equation A4 follows from the equation $E[\ell_\lambda\{\psi, \lambda(\psi, \phi)\}; \hat{\psi}, \phi] = 0$ that defines ϕ. Equation A5 follows from a Taylor series expansion since $\hat{\phi}_\psi = \hat{\lambda} + O(n^{-1/2})O(|\psi - \hat{\psi}|)$. Let $K(\psi) = \{-\ell_{\lambda\lambda}(\psi, \hat{\lambda}_\psi) + \ell_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})\}/\sqrt{n}$; then A6 follows from the expansion

$$\begin{aligned} -\ell_{\lambda\lambda}(\psi, \hat{\lambda}_\psi) - E\{-\ell_{\lambda\lambda}(\psi, \hat{\lambda}_\psi); \hat{\psi}, \hat{\lambda}\} &= -\ell_{\lambda\lambda}(\hat{\psi}, \hat{\lambda}) - E\{-\ell_{\lambda\lambda}(\hat{\psi}, \hat{\lambda}); \hat{\psi}, \hat{\lambda}\} \\ &\quad + \sqrt{n}[K(\psi) - E\{K(\psi); \hat{\psi}, \hat{\lambda}\}] + O(1), \end{aligned}$$

since $-\ell_{\lambda\lambda}(\hat{\psi}, \hat{\lambda}) - E\{-\ell_{\lambda\lambda}(\hat{\psi}, \hat{\lambda}); \hat{\psi}, \hat{\lambda}\}$ depends only on the data.

For a full-rank exponential family model, $\hat{\phi}_\psi = \hat{\lambda}$ so that $I(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\phi}_\psi) = I(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\lambda})$ and, by properties of exponential family models, $E\{-\ell_{\lambda\lambda}(\psi, \hat{\lambda}_\psi); \hat{\psi}, \hat{\lambda}\} = \hat{j}_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)$. Thus, A2 holds exactly.

References

Abramowitz, M. & Stegun, I. A. (1965). *Handbook of Mathematical Functions*. New York: Dover.

Barnard, G. A., Jenkins, G. M. & Winsten, C. B. (1962). Likelihood inference and time series (with Discussion). *J. R. Statist. Soc.* A **125**, 321–75.

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70**, 343–65.

Barndorff-Nielsen, O. E. (1994). Adjusted versions of profile likelihood and directed likelihood, and extended likelihood. *J. R. Statist. Soc.* B **56**, 125–40.

Barndorff-Nielsen, O. E. (1995). Stable and invariant adjusted profile likelihood and directed likelihood for curved exponential models. *Biometrika* **82**, 489–500.

Barndorff-Nielsen, O. E. & Cox, D. R. (1994). *Inference and Asymptotics*. London: Chapman and Hall.

Berger, J. O. & Bernardo, J. M. (1989). Estimating a product of normal means: Bayesian analysis with reference priors. *J. Am. Statist. Assoc.* **84**, 200–7.

Berger, J. O. & Bernardo, J. M. (1992). On the development of reference priors. In *Bayesian Statistics 4*, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, pp. 35–60. Oxford: Oxford University Press.

Berger, J. O., Liseo, B. & Wolpert, R. (1999). Integrated likelihood functions for eliminating nuisance parameters (with Discussion). *Statist. Sci.* **14**, 1–28.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (with Discussion). *J. R. Statist. Soc.* B **41**, 113–47.

COX, D. R. & REID, N. (1987). Parameter orthogonality and approximate conditional inference (with Discussion). *J. R. Statist. Soc.* B **49**, 1–39.
DICICCIO, T. J., MARTIN, M. A., STERN, S. E. & YOUNG, G. A. (1996). Information bias and adjusted profile likelihoods. *J. R. Statist. Soc.* B **58**, 189–203.
EVANS, M. & SWARTZ, T. (2000). *Approximating Integrals Via Monte Carlo and Deterministic Methods*. Oxford: Oxford University Press.
FERGUSON, H., COX, D. R. & REID, N. (1991). Estimating equations from modified profile likelihood. In *Estimating Functions*, Ed. V. P. Godambe, pp. 279–94. Oxford: Oxford University Press.
FRASER, D. A. S. (2003). Likelihood for component parameters. *Biometrika* **90**, 327–40.
FRASER, D. A. S. & REID, N. (1989). Adjustments to profile likelihood. *Biometrika* **76**, 477–88.
JEFFREYS, H. (1983). *Theory of Probability*, 3rd ed. Oxford: Oxford University Press.
KALBFLEISCH, J. D. & SPROTT, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters (with Discussion). *J. R. Statist. Soc.* B **32**, 175–208.
KALBFLEISCH, J. D. & SPROTT, D. A. (1973). Marginal and conditional likelihoods. *Sankhya* A **35**, 311–28.
KASS, R. E. & WASSERMAN, L. (1996). Formal rules for selecting prior distributions. *J. Am. Statist. Assoc.* **91**, 1343–70.
LISEO, B. (1993). Elimination of nuisance parameters with reference priors. *Biometrika* **80**, 295–304.
MARTZ, H. F. & WALLER, R. A. (1982). *Bayesian Reliability Analysis*. New York: Wiley.
MCCULLAGH, P. & TIBSHIRANI, R. (1990). A simple method for the adjustment of profile likelihoods. *J. R. Statist. Soc.* B **52**, 325–44.
PACE, L. & SALVAN, A. (1997). *Principles of Statistical Inference*. Singapore: World Scientific.
SEVERINI, T. A. (1998a). An approximation to the modified profile likelihood function. *Biometrika* **85**, 403–11.
SEVERINI, T. A. (1998b). Likelihood functions for the elimination of nuisance parameters. *Biometrika* **85**, 507–22.
SEVERINI, T. A. (2000). *Likelihood Methods in Statistics*. Oxford: Oxford University Press.
SKOVGAARD, I. M. (1996). An explicit large-deviation approximation to one-parameter tests. *Bernoulli* **2**, 145–65.
SWEETING, T. J. (1987). Discussion of the paper by D. R. Cox and N. Reid. *J. R. Statist. Soc.* B **49**, 20–22.

[*Received November* 2005. *Revised November* 2006]