Statistica Sinica **21** (2011), 433-447

FREQUENCY PROPERTIES OF INFERENCES BASED ON AN INTEGRATED LIKELIHOOD FUNCTION

Thomas A. Severini

Northwestern University

Abstract: One approach to likelihood inference for a parameter of interest in the presence of a nuisance parameter is to use an integrated likelihood in which the nuisance parameter is eliminated from the likelihood by integrating with respect to a prior density. In this paper, the frequency properties of point estimators and interval estimators based on an integrated likelihood function are considered. These results are used to study the problem of choosing the prior density so that the resulting integrated likelihood function is useful for non-Bayesian likelihood inference.

Key words and phrases: Confidence intervals, local power, nuisance parameters, point estimation.

1. Introduction

Consider a model with parameter $\theta \in \Theta$, and suppose that θ may be written $\theta = (\psi, \lambda)$ where ψ is a real-valued parameter of interest and λ is a nusiance parameter. We assume, without loss of generality, that ψ and λ are orthogonal parameters in the sense that corresponding off-diagonal elements of the Fisher information matrix are 0; see, e.g., Cox and Reid (1987). Let $L(\psi, \lambda)$ denote the likelihood function corresponding to a particular set of data and consider likelihood inference for ψ.

In models without a nuisance parameter, inference can be based directly on the likelihood function; 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 instance, one commonly-used pseudolikelihood is the profile likelihood function, given by

$$L_p(\psi) = \sup_{\lambda \in \Lambda} L(\psi, \lambda);$$

here Λ denotes the space of possible λ. Other pseudolikelihood functions include marginal and conditional likelihoods, although these require specific model structures, and modified versions of the profile likelihood; see e.g., Barndorff-Nielsen (1983, 1994), Barndorff-Nielsen and Cox (1994, Chap. 8) Cox and Reid (1987),

Fraser (2003), Fraser and Reid (1989), Kalbfleisch and Sprott (1970, 1973), McCullagh and Tibshirani (1990) and Severini (2000, Chap. 9) for discussion of various approaches to likelihood inference in the presence of a nuisance parameter.

An alternative approach is to eliminate λ in the likelihood function by integrating with respect to a nonnegative weight function $\pi(\lambda|\psi)$ on Λ. We refer to $\pi(\lambda|\psi)$ as the prior density for λ given ψ even though, for our purposes, it is not necessary that π be a genuine density function; also, it should be understood that the prior density for λ is a conditional density given ψ and, hence, it may depend on ψ. Then the integrated likelihood function with respect to π is given by

$$\int_{\Lambda} L(\psi, \lambda)\pi(\lambda|\psi)d\lambda; \tag{1.1}$$

see, for example, Berger, Liseo, and Wolpert (1999), Kalbfleisch and Sprott (1970), Liseo (1993), and Severini (2007) for further discussion of integrated likelihoods.

Thus, unlike pseudolikelihoods based on the profile likelihood, integrated likelihoods are based on averaging rather than maximization and, unlike marginal and conditional likelihoods, integrated likelihoods are always available. Of course, integrated likelihood functions have the drawback that the prior density must be chosen.

One approach to selecting the prior density is to attempt to construct an integrated likelihood function that has properties similar to those of a genuine likelihood. Two such properties are score-unbiasedness and information-unbiasedness; see, e.g., DiCiccio et al. (1996), and Lindsay (1982). A pseudolikelihood for ψ is score-unbiased if its log-derivative with respect to ψ has mean 0; it is information-unbiased if the second moment of its first log-derivative plus the first moment of its second log-derivative is 0. It has been shown that, if π is chosen so that the integrated likelihood function $\bar{L}(\psi)$ is approximately score unbiased, then $\bar{L}(\psi)$ is approximately equal to the Cox-Reid adjusted profile likelihood. If π is chosen so that the integrated likelihood is approximately score-unbiased and approximately information-unbiased, then $\bar{L}(\psi)$ is approximately equal to the modified profile likelihood; these results are discussed in more detail in Section 3.

However, properties such as score bias are not used directly in the construction of statistical procedures. Thus, it is possible that an alternative integrated likelihood, not approximately score-unbiased or approximately information-unbiased, may yield better statistical properties.

The goal of this paper is to consider the asymptotic frequency properties of point estimators, and interval estimators based on an integrated likelihood function. These results are used to study the problem of choosing the prior density so

that the resulting integrated likelihood function is useful for non-Bayesian likelihood inference. It is shown that, at least in some cases, it is possible to construct a prior density that yields procedures that are asymptotically superior to those based on a score-unbiased integrated likelihood.

2. Notation and Some Preliminary Results

2.1. Notation and assumptions

Let $L(\psi, \lambda)$ denote the likelihood function for the model and let $\ell(\psi, \lambda) = \log L(\psi, \lambda)$ denote the log-likelihood. Assume that L is based on n independent, identically distributed observations, that the model is regular in the sense $\ell(\psi, \lambda)$ can be approximated by a polynomial, and that integration and differentiation can be interchanged; see Severini (2000, Sec. 3.4) for further discussion.

Derivatives of $\ell(\psi, \lambda)$ with respect to (ψ, λ) will be denoted by subscripts so that, for example,

$$\ell_\psi(\psi, \lambda) = \frac{\partial}{\partial \psi}\ell(\psi, \lambda), \quad \ell_\lambda(\psi, \lambda) = \frac{\partial}{\partial \lambda}\ell(\psi, \lambda), \quad \ell_{\psi\lambda}(\psi, \lambda) = \frac{\partial^2}{\partial \psi \partial \lambda^{\mathrm{T}}}\ell(\psi, \lambda).$$

Here $\ell_\psi(\psi, \lambda)$ is a scalar, $\ell_\lambda(\psi, \lambda)$ is a $d \times 1$ vector, and $\ell_{\psi\lambda}(\psi, \lambda)$ is $1 \times d$ vector, where d denotes the dimension of λ.

Let $\ell^{(1)}$ denote the log-likelihood for a single observation. Expected values of derivatives of $\ell^{(1)}$ are denoted by μ, with the subscripts of μ indicating the derivatives under consideration. For example,

$$\mu_{\psi\psi}(\psi, \lambda) = \mathrm{E}\{\ell^{(1)}_{\psi\psi}(\psi, \lambda); \psi, \lambda\}, \quad \mu_{\psi,\lambda}(\psi, \lambda) = \mathrm{E}\{\ell^{(1)}_{\psi}(\psi, \lambda)\ell^{(1)}_{\lambda}(\psi, \lambda)^{\mathrm{T}}; \psi, \lambda\},$$

$$\mu_{\psi\lambda,\lambda}(\psi, \lambda) = \mathrm{E}\{\ell^{(1)}_{\psi\lambda}(\psi, \lambda)\ell^{(1)}_{\lambda}(\psi, \lambda)^{\mathrm{T}}; \psi, \lambda\},$$

and so on. Note that $\mu_{\psi\psi}$ is a scalar, $\mu_{\psi,\lambda}$ is a $1 \times d$ vector, and $\mu_{\psi\lambda,\lambda}$ is a $d \times d$ matrix; also note that, in some cases, the dependence of these quantities on (ψ, λ) is often suppressed.

Let

$$\begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix} = \begin{pmatrix} \mu_{\psi,\psi} & \mu_{\psi,\lambda} \\ \mu_{\lambda,\psi} & \mu_{\lambda,\lambda} \end{pmatrix} = - \begin{pmatrix} \mu_{\psi\psi} & \mu_{\psi\lambda} \\ \mu_{\lambda\psi} & \mu_{\lambda\lambda} \end{pmatrix}$$

denote the expected information matrix for a single observation. By the assumed orthogonality of ψ and λ, $i_{\psi\lambda} = i^{\mathrm{T}}_{\lambda\psi} = 0$.

2.2. Prior densities

Consider the integrated likelihood function

$$\bar{L}(\psi) = \int_\Lambda L(\psi, \lambda)\pi(\lambda|\psi)d\lambda, \tag{2.1}$$

where $\pi(\lambda|\psi)$ is a given prior density. When considering the properties of different prior densities, we must deal with the issue of standardization. Assume that either

$$\int_{\Lambda} \pi(\lambda|\psi)d\lambda \tag{2.2}$$

has the same finite value for each ψ or, (2.2) is infinite for each ψ and $\pi(\lambda|\psi)$ has been normalized using the approach described in Berger, Liseo, and Wolpert (1999).

This approach is based on a sequence of nested subsets $\Omega_1, \Omega_2, \ldots$ of $\Psi \times \Lambda$, where Ψ denotes the space of ψ. Assume that Ω_m increases to $\Psi \times \Lambda$ as $m \to \infty$, and let $\Lambda_m = \{\lambda : (\psi, \lambda) \in \Omega_m\}$. Define

$$K_m(\psi) = \int_{\Lambda_m} \pi(\lambda|\psi)d\psi, \quad m = 1, 2, \ldots.$$

Assume that, for any ψ in the interior of Ψ,

$$\lim_{m\to\infty} \frac{K_m(\psi)}{K_m(\psi_0)}$$

exists, does not depend on ψ, and depends on ψ_0 only through a proportionality constant. It is important to note that different sequences $\Lambda_1, \Lambda_2, \ldots$ will lead to different normalization factors and, hence, different intergrated likelihoods. See Berger, Liseo, and Wolpert (1999) for further details on this type of normalization.

2.3. Laplace approximation

Consider the integrated likelihood $\bar{L}(\psi)$ based on a prior $\pi(\lambda|\psi)$. Using a Laplace approximation (see, e.g., Evans and Swartz (2000, Chap. 4)) for the integral in (2.1), it follows that

$$\bar{L}(\psi) = \int_{\Lambda} (\psi, \lambda)\pi(\lambda|\psi)d\lambda \ = c_0 \ L(\psi, \hat{\lambda}_\psi)|-\ell_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{-1/2}\pi(\hat{\lambda}_\psi|\psi)\{1 + D_n(\psi)\}, \tag{2.3}$$

where $\hat{\lambda}_\psi$ denotes the maximum likelihood estimate of λ for fixed ψ; here $D_n(\psi) = O(n^{-1})$ for any fixed ψ, and c_0 does not depend on ψ. Let $\hat{\psi}$ denote the maximum likelihood estimator of ψ. For $\psi = \hat{\psi} + O(n^{-1/2})$, $D_n(\psi) = D_n(\hat{\psi})[1 + O(n^{-1/2})]$ so that, by modifying the definition of c_0, the expansion given in (2.3) holds with error $O(n^{-3/2})$.

Note that $\bar{L}(\psi)$ can be written as

$$\bar{L}(\psi) = L_A(\psi)\pi(\hat{\lambda}_\psi|\psi)[1 + D_n(\psi)],$$

where L_A denotes the Cox-Reid adjusted profile likelihood (Cox and Reid (1987)), given by

$$L_A(\psi) = L(\psi, \hat{\lambda}_\psi)|-\ell_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{-1/2};$$

here, as elsewhere in the paper, multiplicative terms not depending on ψ have been ignored.

3. Score-Unbiased Integrated Likelihoods

In general, an integrated likelihood function is not a likelihood function, in the sense that it is not based on a marginal or conditional density function. Thus, an integrated likelihood does not necessarily have the frequency properties of a genuine likelihood function. Important properties of this type are the first two Bartlett identities. If $M(\psi)$ is a genuine likelihood function for ψ and $m(\psi) = \log M(\psi)$, the first Bartlett identity states that $\mathrm{E}\{m'(\psi); \theta\} = 0$, known as score unbiasedness. The second Bartlett identity states that

$$\mathrm{E}\{m''(\psi) + m'(\psi)m'(\psi)^{\mathrm{T}}; \theta\} = 0,$$

known as information unbiasedness.

If $\bar{L}(\psi)$ is an integrated likelihood function and $\bar{\ell}(\psi) = \log \bar{L}(\psi)$ then, in general, $\mathrm{E}\{\bar{\ell}'(\psi); \theta\}$ and $\mathrm{E}\{\bar{\ell}'(\psi) + \bar{\ell}'(\psi)\bar{\ell}'(\psi)^{\mathrm{T}}; \theta\}$ are both $O(1)$ as $n \to \infty$ (Severini (1998)).

Consider an integrated likelihood of the form

$$\bar{L}(\psi) = \int_\Lambda L(\psi, \lambda)\pi(\lambda|\psi)d\lambda$$

and let $\bar{\ell}(\psi) = \log \bar{L}(\psi)$. If $\pi(\lambda|\psi)$ does not depend on ψ then, ignoring constants not depending on ψ, $\bar{\ell}(\psi) = \ell_A(\psi) + O(n^{-1})$, where $\ell_A(\psi)$ denotes the Cox-Reid adjusted log-likelihood; see Sweeting (1987) and Severini (2007) for further discussion.

Using this result, it follows that if $\pi(\lambda|\psi)$ does not depend on ψ, then $\bar{L}(\psi)$ is score-unbiased to $O(n^{-1})$: $\mathrm{E}\{\bar{\ell}'(\psi); \theta\} = O(n^{-1})$; see Ferguson, Reid, and Cox (1991). Furthermore, suppose that $\bar{L}_0$ is an integrated likelihood such that

$$\mathrm{E}\{\bar{\ell}_0'(\psi); \psi, \lambda\} = O(n^{-1}). \tag{3.1}$$

By (2.3), for $\psi = \hat{\psi} + O(n^{-1/2})$, $\bar{\ell}_0(\psi) = \ell_A(\psi) + h(\psi, \hat{\lambda}_\psi) + O(n^{-3/2})$, where $h(\psi, \lambda) = \log \pi(\lambda|\psi)$. Let h_ψ, h_λ denote the derivatives of h with respect to ψ and λ, respectively, and let $\hat{\lambda}'_\psi = d\hat{\lambda}_\psi/d\psi$. Then

$$\frac{d}{d\psi}h(\psi, \hat{\lambda}_\psi) = h_\lambda(\psi, \hat{\lambda}_\psi)\hat{\lambda}'_\psi + h_\psi(\psi, \hat{\lambda}_\psi).$$

For $\psi = \hat{\psi} + O(n^{-1/2})$, $\hat{\lambda}'_\psi = O(n^{-1/2})$ and $\hat{\lambda}_\psi = \lambda + O(n^{-1/2})$. Thus, for $\psi = \hat{\psi} + O(n^{-1/2})$,

$$\frac{d}{d\psi} h_\psi(\psi, \hat{\lambda}_\psi) = h_\psi(\psi, \lambda) + O_p(n^{-1/2})$$

and, hence,

$$\bar{\ell}'_0(\psi) = \ell'_A(\psi) + h_\psi(\psi, \lambda) + O(n^{-1}).$$

It follows that if (3.1) holds, then $h_\psi(\psi, \lambda) = 0$ for all λ and ψ. Hence, under (3.1), for $\psi = \hat{\psi} + O(n^{-1/2})$, $h(\psi, \hat{\lambda}_\psi) = h(\hat{\psi}, \hat{\lambda}) + O_p(n^{-1})$, so that $\pi(\hat{\lambda}_\psi|\psi) = \pi(\hat{\lambda}|\hat{\psi})\{1 + O(n^{-1})\}$.

It now follows from (2.3) that $\bar{L}_0(\psi)$ can be approximated by $L_A(\psi)$, with error $O(n^{-1})$ for $\psi = \hat{\psi} + O(n^{-1/2})$; that is, any integrated likelihood function $\bar{L}_0(\psi)$ that is approximately score unbiased is approximately equal to the Cox-Reid adjusted profile likelihood. In particular, the modified profile likelihood and approximations to the modified profile likelihood agree with L_A to order $O(n^{-1})$.

4. Frequency Properties of Inferences Based on an Integrated Likelihood

4.1. Introduction

The results described in the previous section show that if the goal is to construct an integrated likelihood function that is approximately score unbiased, then that integrated likelihood can be approximated by the Cox-Reid adjusted profile likelihood.

However, properties such as score bias are not used directly in the construction of statistical procedures. Thus, it is possible that an alternative integrated likelihood, that is not score unbiased, may yield statistical procedures that are superior to those based on the Cox-Reid adjusted profile likelihood.

In this section, the frequency properties of procedures based on an integrated likelihood are considered. In particular, the dependence of the frequency properties on the prior density are considered, and the possibility of improving on the properties of those procedures based on $L_A(\psi)$ is explored. Note that, although the comparisons considered here are described in terms of L_A, the same results hold for the modified profile likelihood as well as approximations to the modified profile likelihood.

We consider two frequency properties of likelihood-based methods: the coverage probability of an integrated-likelihood ratio confidence interval and the mean squared error of the maximum integrated likelihood estimator. In each case, asymptotic expansions of the relevant property are presented and the implications of those results for the selection of $\pi(\lambda|\psi)$ are considered.

Recall that an integrated likelihood function based on a prior density $\pi(\lambda|\psi)$ can be approximated by $L_A(\psi)\pi(\hat{\lambda}_\psi|\psi)$. It is shown that the frequency properties under consideration depend on $\pi(\lambda|\psi)$ through $h_\psi(\psi,\lambda) = \partial h(\psi,\lambda)/\partial\psi$, where $h(\psi,\lambda) = \log\pi(\lambda|\psi)$.

4.2. Integrated-likelihood ratio confidence intervals

Let $\bar{W}(\psi) = 2[\bar{\ell}(\bar{\psi}) - \bar{\ell}(\psi)]$, where $\bar{\ell}(\psi) = \log\bar{L}(\psi)$ and $\bar{\psi}$ is the value of ψ that maximizes $\bar{L}(\psi)$. If $F(\psi,\lambda) = n[\mathrm{E}\{\bar{W}(\psi);\psi,\lambda\} - 1]$, then $F(\psi,\lambda) = O(1)$ and $1 + F(\psi,\lambda)/n$ is the Bartlett correction factor for the statistic $\bar{W}(\psi)$; see DiCiccio and Stern (1994) for discussion of the result that $\bar{W}(\psi)$ is Bartlett-correctable. If $\bar{W}_c(\psi) = \bar{W}(\psi)/[1 + F(\hat{\psi},\hat{\lambda})]$, then under the distribution with parameter (ψ,λ), $\bar{W}_c(\psi)$ has a chi-squared distribution with error $o(n^{-1})$.

The $(1-\alpha)\times100\%$ integrated likelihood ratio confidence region based on $\bar{L}(\psi)$ consists of those values of ψ for which $\bar{W}(\psi) \le \chi^2{}_1(\alpha)$, where $\chi^2{}_1(\alpha)$ denotes the $1-\alpha$-quantile of the chi-squared distribution with one degree-of-freedom. Although this confidence region is not necessarily an interval, Mukerjee and Reid (1999) show that it can be approximated by an interval that has coverage probability $\alpha + o(n^{-1})$.

Let $\bar{C}$ denote the length of the confidence interval based on the integrated likelihood and let C_A denote the length of the confidence interval based on L_A. Then, using the results of Mukerjee and Reid (1999),

$$\mathrm{E}\{\sqrt{n}\bar{C}\} = \mathrm{E}\{\sqrt{n}C_A\} + \frac{1}{n}\frac{z_{\alpha/2}}{\sqrt{i_{\psi\psi}(\psi,\lambda)}}\Delta_L + o(\frac{1}{n})$$
$$\Delta_L(\psi,\lambda) = 2\frac{\partial}{\partial\psi}\frac{h_\psi(\psi,\lambda)}{i_{\psi\psi}(\psi,\lambda)} + \frac{h_\psi(\psi,\lambda)^2}{i_{\psi\psi}(\psi,\lambda)};$$

here z_α denotes the $1-\alpha$ quantile of the standard normal distribution. Thus, the expected length of the integrated likelihood confidence interval tends to be small whenever $h_\psi(\psi,\lambda)/i_{\psi\psi}(\psi,\lambda)$ is a decreasing function of ψ. If $\partial[h_\psi(\psi,\lambda)/i_{\psi\psi}(\psi,\lambda)]/\partial\psi$ is sufficiently negative, then the expected length of the integrated likelihood confidence interval is less than that of the confidence interval based on L_A.

4.3. Mean squared error of the maximum integrated likelihood estimator

As above, let $\bar{\psi}$ denote the value of ψ that maximizes $\bar{L}(\psi)$. Then $\bar{\psi}$ can be used as a point estimator of ψ. Let $\hat{\psi}_A$ denote the value of ψ that maximizes $L_A(\psi)$. The results of Mukerjee and Reid (1999) can be used to show that

$$n\mathrm{E}\{(\bar{\psi}-\psi)^2;\psi,\lambda\} = n\mathrm{E}\{(\hat{\psi}_A-\psi)^2;\psi,\lambda\} + \frac{1}{n}\Delta_M(\psi,\lambda) + o(\frac{1}{n}),$$

where

$$\Delta_M(\psi,\lambda) = \frac{1}{i_{\psi\psi}(\psi,\lambda)}\{\Delta_L(\psi,\lambda) + \frac{1}{i_{\psi\psi}^2}[2\mu_{\psi,\psi\psi}(\psi,\lambda) + \mu_{\psi\psi\psi}(\psi,\lambda)]h_\psi(\psi,\lambda)\}.$$

Thus if $\Delta_M(\psi,\lambda) < 0$ then $\bar{\psi}$ is preferable to $\hat{\psi}_A$ as an estimator of ψ when (ψ,λ) is the true parameter. As with the case of the expected confidence interval length, the mean squared error of the maximum integrated likelihood estimator tends to be small whenever $h_\psi(\psi,\lambda)/i_{\psi\psi}(\psi,\lambda)$ is a decreasing function of ψ. If $\partial[h_\psi(\psi,\lambda)/i_{\psi\psi}(\psi,\lambda)]/\partial\psi$ is sufficiently negative, then the mean squared error of the maximum integrated likelihood estimator is less than that of the maximizer of L_A.

It is important to note that the type of comparison implicit in the use of Δ_M is not uniform in (ψ,λ), so that this type of asymptotic mean squared comparison is different than the type of comparisons used when considering the admissibility of estimators. The estimator $\hat{\psi}_A$ is inadmissible if $\mathrm{E}\{(\hat{\psi}_A - \psi)^2; \psi, \lambda\} \geq \mathrm{E}\{(\bar{\psi} - \psi)^2; \psi, \lambda\}$ for all ψ, λ, with strict inequality for some (ψ,λ). If $\Delta_M(\psi,\lambda) < 0$ then, for sufficiently large n,

$$\mathrm{E}\{(\hat{\psi}_A - \psi)^2; \psi, \lambda\} < \mathrm{E}\{(\bar{\psi} - \psi)^2; \psi, \lambda\}; \tag{4.1}$$

however, how large n needs to be for (4.1) to hold may depend on (ψ,λ). Thus, it is possible to have $\Delta_M(\psi,\lambda) < 0$ for all ψ, λ even when $\hat{\psi}_A$ is an admissible estimator.

4.4. The effect of reparameterization

We now consider the effect of reparameterization on the quantities Δ_L and Δ_M. Two types of reparameterization are of interest. One is reparametrization of the nuisance parameter in terms of $\phi = g(\lambda)$ for some smooth function g. All of the properties discussed above are invariant with respect to this type of reparameterization.

A second type of reparameterization is reparameterization of the parameter of interest. Let $\eta = q(\psi)$, where q is a smooth, one-to-one function on the space of possible ψ, and let $Q = q^{-1}$. Consider the analysis of confidence intervals, hypothesis tests, and point estimators of η and let $\tilde{\Delta}_L(\eta,\lambda)$ and $\tilde{\Delta}_M(\eta,\lambda)$ denote Δ_L and Δ_M, respectively, based on this new parameterization. Let $\tilde{\pi}(\lambda|\eta)$ denote the prior density for constructing the integrated likelihood for η, and let $\pi(\lambda|\psi)$ denote the corresponding prior in the original parameterization. Then $\tilde{\pi}(\lambda|\eta) = \pi(\lambda|Q(\eta))$ and, hence,

$$\tilde{h}_\eta(\eta,\lambda) = \frac{\partial}{\partial\eta}\log\tilde{\pi}(\lambda|\eta) = h_\psi(Q(\eta),\lambda)Q'(\eta).$$

Using standard properties of log-likelihood derivatives under reparameterization, it follows that

$$\tilde{\Delta}_L(\eta,\lambda) = \Delta_L(Q(\eta),\lambda) - \frac{h_\psi(Q(\eta),\lambda)Q''(\eta)}{i_{\psi\psi}(Q(\eta),\lambda)Q'(\eta)^2},$$
$$\tilde{\Delta}_M(\eta,\lambda) = \frac{\Delta_M(Q(\eta),\lambda)}{Q'(\eta)^2} - 2\frac{Q''(\eta)}{i_{\psi\psi}(Q(\eta),\lambda)^2 Q'(\eta)^4} h_\psi(Q(\eta),\lambda).$$

Thus, comparisons based on confidence interval length and the mean squared error of estimators depend on the parameterization used. Therefore, when considering specific examples, it is important to keep in mind that different conclusions might be reached if the parameter of interest is reparameterized.

5. Examples

5.1. Combining the results from two apparently unrelated experiments

The purpose of this example is to show how using a prior density for λ that depends on ψ might lead to improved inferences for ψ, even in cases in which ψ and λ are not statistically related. In this example, ψ and λ are strongly unrelated, in the sense that the likelihood function for (ψ,λ) factors into a function of ψ times a function of λ.

Let $X_1,\ldots,X_n$ and $Y_1,\ldots,Y_n$ denote independent random variables such that for each $j=1,\ldots,n$, X_j has a normal distribution with mean ψ and standard deviation 1, and Y_j has a normal distribution with mean λ and standard deviation 1. Here ψ and λ both take values in $\Re$. Clearly, ψ and λ are orthogonal parameters and, in fact, there is no statistical connection between them; the integrated likelihood for ψ based on a prior for λ that does not depend on ψ is simply the likelihood for ψ using only $X_1,\ldots,X_n$. However, if the X_j and Y_j are measurements on similar quantities, it may make sense to use $Y_1,\ldots,Y_n$ to improve estimation of ψ, as is done in the case of empirical Bayes estimation. This is true even though $(X_1,\ldots,X_n)$ and $(Y_1,\ldots,Y_n)$ are independent.

The prior density $\pi(\lambda|\psi)$ implies a stochastic relationship between ψ and λ that can be used to extract information regarding ψ from the distribution of $Y_1,\ldots,Y_n$. For instance, consider the prior density

$$\pi(\lambda|\psi) \propto \exp\{-\frac{\omega}{2}(\lambda-\psi)^2\}, \quad -\infty<\lambda<\infty, \tag{5.1}$$

where $\omega>0$ is known constant; the selection of ω is considered below. Thus, given ψ, we assume that $\lambda-\psi$ has a normal distribution with mean 0 and variance $1/\omega$, and use of this prior can be interpreted as an assumption that there is some relationship between ψ and λ.

It is straightforward to show that for this model

$$\Delta_L(\psi,\lambda) = \Delta_M(\psi,\lambda) = 2\frac{d}{d\psi}h_\psi(\psi,\lambda) + h_\psi(\psi,\lambda)^2.$$

For the prior given by (5.1), $h_\psi(\psi,\lambda) = -\frac{\omega}{2}(\lambda-\psi)^2$ and, hence,

$$\Delta_L(\psi,\lambda) = \omega^2\{(\lambda-\psi)^2 - \frac{2}{\omega}\}. \tag{5.2}$$

Thus, confidence intervals and point estimators for ψ based on the integrated likelihood using prior (5.1) are superior to those based on the likelihood function for ψ based on $X_1,\ldots,X_n$ for all ψ,λ such that

$$(\lambda-\psi)^2 < \frac{2}{\omega}. \tag{5.3}$$

The quantity $\Delta_L(\psi,\lambda)$ is negative whenever λ and ψ are sufficiently close, with the parameter ω governing how close λ and ψ must be for the integrated likelihood approach to be beneficial. Thus, the choice of ω depends on our assumptions regarding the relationship between ψ and λ. For instance, if ω is chosen to be very large then, effectively, $\lambda = \psi$ so that $Y_1,\ldots,Y_n$ contain considerable information about ψ. On the other hand, if the relationship between ψ and λ is considered to be very weak, a small value of ω is appropriate. Note that the set of parameter values for which (5.3) is satisfied can be increased by choosing ω to be small; however, (5.2) shows that a small value of ω decreases the magnitude of $\Delta(\psi,\lambda)$. That is, a stronger assumption about the relationship between λ and ψ potentially results in improved inferences for ψ, but the set of parameter values for which this stronger assumption is valid is relatively small.

The prior density (5.1) states that $\lambda-\psi$ has a normal distribution with variance $1/\omega$. Thus, (5.3) can be written

$$\frac{(\lambda-\psi)^2}{\mathrm{E}_\pi[(\lambda-\psi)^2|\psi]} < 2, \tag{5.4}$$

where E_π denotes expectation with respect to the prior (5.1). This gives a useful interpretation of the results described above. If λ and ψ satisfy the stochastic relationship implied by the prior (5.1), in the sense that (5.4) holds, then the integrated likelihood function yields inferences that are superior to those based on the likelihood based on $X_1,\ldots,X_n$.

The integrated likelihood based on (5.1) is given by

$$\bar{L}(\psi) = \exp\{-\frac{n}{2}(\psi-\bar{X})^2\}\exp\{-\frac{1}{2}\frac{n\omega}{n+\omega}(\psi-\bar{Y})^2\}.$$

Thus, in terms of the likelihood functions, the difference between the integrated likelihood based on (5.1) and the likelihood based only on $X_1, \ldots, X_n$, is equivalent to the observation of an additional random variable that has a normal distribution with mean ψ and variance $1/\omega + 1/n$. However, when considering frequency properties of the resulting inferences, it is important to keep in mind that the mean of this random variable is λ, not ψ.

5.2. Ratio of normal means

Let $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ denote independent random variables such that $X_1, \ldots, X_n$ each have a normal distribution with mean μ_X and $Y_1, \ldots, Y_n$ each have a normal distribution with mean μ_Y; assume that $\mu_X > 0$ and $\mu_Y > 0$. The parameter of interest is the ratio of the means, $\psi = \mu_X/\mu_Y$, and take $\lambda = \sqrt{(\mu_X^2 + \mu_Y^2)}$ as the nuisance parameter; then ψ and λ are orthogonal.

Consider construction of a confidence interval for ψ. It is straightforward to show that

$$\Delta_L(\psi, \lambda) = 4\frac{\partial}{\partial\lambda}\psi^2 h_\psi(\psi, \lambda) + 2\psi^2 h_\psi(\psi, \lambda)^2.$$

The average expected information in the model with ψ known is given by $i_{\lambda\lambda}(\psi, \lambda) = 1$. Thus, if a reference prior is normalized using rectangles defined in terms of μ_X and μ_Y, the reference prior is $(1 + \psi^2)^{-1/2}$. This leads to the integrated likelihood function given by $L_I(\psi) = L_p(\psi)/\sqrt{(1+\psi^2)}$; here L_p denotes the profile likelihood,

$$L_p(\psi) = \exp\{-\frac{n}{2}\frac{(\bar{x} - \psi\bar{y})^2}{1 + \psi^2}\},$$

where $\bar{x}$ and $\bar{y}$ denote the sample means. The profile likelihood can also be obtained as a integrated likelihood with respect to a uniform prior for λ; thus, L_p agrees with L_A, ignoring terms of order $O(n^{-1})$. Note that the profile likelihood has the property that it does not approach 0 as $\psi \to \infty$.

For this prior density,

$$\Delta_L(\psi, \lambda) = -\frac{5\psi^2 + 2}{\lambda^2}.$$

Thus, in terms of confidence interval length, the integrated likelihood $L_I(\psi)$ is superior to the profile likelihood.

Liseo (1993) proposes the use of $L_R(\psi) = L_p(\psi)/(1 + \psi^2)$ for likelihood inference for ψ. The use of L_R is equivalent to the use of the prior $(1 + \psi^2)^{-1}$ for λ; for this prior

$$\Delta_L(\psi, \lambda) = -\frac{8(\psi^2 + 1)}{\lambda^2}.$$

It follows that confidence intervals based on L_R are asymptotically shorter than those based on either L_I or L_p; recall that all confidence intervals being compared have the same coverage probability, neglecting terms of order $o(n^{-1})$. This is in agreement with the conclusions in Liseo (1993), where inferences based on L_R were compared with those based on L_p and related pseudolikelihood functions.

5.3. Ratio of exponential means

Let $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ denote independent random variables such that $X_1, \ldots, X_n$ are each exponentially distributed with mean $\sqrt{\psi}/\lambda$, and $Y_1, \ldots, Y_n$ are each exponentially distributed with mean $1/[\lambda\sqrt{\psi}]$. The parameter of interest ψ is $\mathrm{E}(X_1)/\mathrm{E}(Y_1)$; the nuisance parameter λ is chosen so that ψ and λ are orthogonal. It is straightforward to show that

$$\Delta_L(\psi, \lambda) = 4\frac{\partial}{\partial\psi}\psi^2 h_\psi(\psi, \lambda) + 2\psi^2 h_\psi(\psi, \lambda)^2,$$
$$\Delta_M(\psi, \lambda) = 2\psi^2[\Delta_L(\psi, \lambda) + 2\psi h_\psi(\psi, \lambda)].$$

Consider a uniform prior for $\mathrm{E}(X_1; \psi, \lambda)^{-1}$. Normalizing such a density over the regions

$$\Lambda_M = \{\lambda > 0 : \mathrm{E}(X_1; \psi, \lambda)^{-1} < M\}, \quad M > 0, \tag{5.5}$$

yields the prior for λ given by

$$\pi(\lambda|\psi) = \frac{1}{\sqrt{\psi}}, \quad \lambda > 0. \tag{5.6}$$

With this choice of prior, it is straightforward to show that

$$\Delta_L(\psi, \lambda) = -\frac{3}{2} \quad \text{and} \quad \Delta_M = -5\psi^2.$$

Thus, interval and point estimates based on the integrated likelihood using prior (5.6) are superior to those based on L_A, to the order considered.

The integrated likelihood based on the prior (5.6) is

$$\bar{L}(\psi) = \psi^{-(n+1)}(\frac{\hat{\psi}}{\psi} + 1)^{-(2n+1)}, \tag{5.7}$$

where $\hat{\psi} = \sum_{j=1}^n X_j / \sum_{j=1}^n Y_j$ denotes the maximum likelihood estimator of ψ. Using λ as the nuisance parameter, the Cox-Reid adjusted profile likelihood is

$$L_A(\psi) = \psi^{-(n+1/2)}(\frac{\hat{\psi}}{\psi} + 1)^{-(2n+1)}.$$

It is straightforward to show that $\bar{L}(\psi)$ is maximized by $\bar{\psi} = n\hat{\psi}/(n+1)$ and that $L_A(\psi)$ is maximized by $\hat{\psi}_A = \hat{\psi}$.

Table 1. Exact confidence interval lengths and mean squared errors in Section 5.3.

	CI Length		MSE	
n	$\bar{L}$	L_A	$\bar{L}$	L_A
5	4.51	5.26	0.653	1.000
10	2.28	2.46	0.242	0.306
20	1.40	1.46	0.109	0.123

For this model, it is possible to calculate the exact mean squared errors of $\bar{\psi}$ and $\hat{\psi}_A$, as well as the exact expected lengths of confidence intervals based on $\bar{L}$ and L_A. These values are given in Table 1 for $n = 5, 10, 20$, coverage probability of 95%, and $\psi = 1$; since ψ is a scale parameter in this model, expected lengths and mean squared errors for other values of ψ can be obtained by multiplying the values in the table by ψ and ψ^2, respectively.

The results in Table 1 show that the conclusions based on the asymptotic expansions considered in Section 4, as reflected in the quantities Δ_L and Δ_M, are valid even in small samples. For instance, compare $\hat{\psi}$ and $\bar{\psi}$ for the case $n = 10$, $\psi = 1$. Since $\Delta_M = -5\psi^2$,

$$n^2[\mathrm{E}\{(\bar{\psi} - \psi)^2; \psi, \lambda\} - E\{(\hat{\psi} - \psi)^2; \psi, \lambda\}] = -5\psi^2 + o(1);$$

hence, for $n = 10$, $\psi = 1$, we expect the mean squared error of $\bar{\psi}$ to be approximately 0.05 less than that of $\hat{\psi}$. According to results in Table 1, the mean squared error of $\bar{\psi}$ is actually 0.064 less than that of $\hat{\psi}$. Consideration of the other values in Table 1 shows that conclusions based on Δ_M and Δ_L are valid for small samples, at least for the cases considered.

6. Discussion

Use of an integrated likelihood function requires selection of the prior density $\pi(\lambda|\psi)$. One approach is to choose the prior so that the integrated likelihood function is approximately score-unbiased. This can be achieved by choosing $\pi(\lambda|\psi)$ so that it does not depend on ψ; recall that ψ and λ are assumed to be orthogonal. The resulting integrated likelihood function is then asymptotically equivalent to the Cox-Reid adjusted profile likelihood. However, the results presented in this paper show that an integrated likelihood function that is not approximately score unbiased may yield statistical procedures with improved performance.

One explanation for this result is based on a generalization of the example in Section 5.1. The prior density places a stochastic constraint on the parameters; such a constraint effectively allows some of the information available for inference for λ to be used for inference about ψ. The integrated likelihood yields improved inferences for ψ when the true parameter values satisfy the stochastic constraint,

in some sense. Thus, the prior $\pi(\lambda|\psi)$ should be chosen so that the constraint implied by the prior is appropriate for the model, and data, under consideration.

Similar conclusions can be reached by analyzing the likelihood function directly. Using a Laplace approximation, the integrated likelihood function based on the prior $\pi(\lambda|\psi)$ satisfies

$$\bar{L}(\psi) = L_A(\psi)\pi(\hat{\lambda}_\psi|\psi)[1 + O(n^{-1})] = L_A(\psi)\pi(\hat{\lambda}|\psi)[1 + O(n^{-1})]$$

for $\psi = \hat{\psi} + O(n^{-1})$. Thus, compared to $L_A(\psi)$, $\bar{L}(\psi)$ includes an additional "observation" $\hat{\lambda}$ with corresponding likelihood contribution $\pi(\hat{\lambda}|\psi)$. Of course the density of $\hat{\lambda}$ is not $\pi(\cdot|\psi)$; the additional likelihood contribution is only useful if $\pi(\hat{\lambda}|\psi)$ is "close to" the true likelihood based on $\hat{\lambda}$, in some sense.

One limitation of the analysis in this paper is it is based on the asymptotic scenario in which the number of nuisance parameters remains fixed as the sample size increases. Thus the results presented here are relevant whenever the number of parameters is small relative to the sample size. In these cases, inferences based on an integrated likelihood are first-order equivalent to inferences based on the profile likelihood, so that the effect of different prior densities can be expected to be relatively minor unless the sample size is quite small.

For cases in which the number of parameters is large relative to the sample size, asymptotic theory in which the number of parameters grows with n, may be more appropriate; see, for example, Barndorff-Nielsen and Cox (1994) and Sartori (2003). It is well-known that, in these cases, standard likelihood methods such as those based on the profile likelihood often perform poorly, and the choice of prior density may have a greater effect on the properties of the resulting integrated likelihood.

Another limitation of the analysis is that it applies only to the case in which ψ and λ are orthogonal. Since ψ is a scalar parameter, this can always be achieved by reparameterization of the nuisance parameter (Cox and Reid (1987)), although solving the necessary differential equations can be difficult.

Acknowledgement

I would like to thank A. Salvan, the Editor, and a referee for several useful comments. This work was supported by the National Science Foundation.

References

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. Roy. Statist. Soc. Ser. B* **56**, 125-140.

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

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

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B* **49**, 1-39.

DiCiccio, T. J., Martin, M. A., Stern, S. E. and Young, G. A. (1996). Information bias and adjusted profile likelihoods. *J. Roy. Statist. Soc. Ser. B* **58**, 189-203.

DiCiccio, T. J. and Stern, S. E. (1994). Constructing approximately standard normal pivots from signed roots of adjusted likelihood ratio statistics. *Scand. J. Statist.* **21**, 447-460.

Evans, M. and Swartz, T. (2000). *Approximating Integrals Via Monte Carlo and Deterministic Methods.* Oxford University Press, Oxford.

Ferguson, H., Reid, N. and Cox, D. R. (1991). Estimating functions from modified profile likelihood. In *Estimating Functions* (Edited by V. P. Godambe), 279-293. Oxford University Press, Oxford.

Fraser, D. A. S. (2003). Likelihood for component parameters. *Biometrika* **90**, 327-340.

Fraser, D. A. S. and Reid, N. (1989). Adjustments to profile likelihood. *Biometrika* **76**, 477-88.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters (with discussion). *J. Roy. Statist. Soc. Ser. B* **32**, 175-208.

Kalbfleisch, J. D. and Sprott, D. A. (1973). Marginal and conditional likelihoods. *Sankhyā A* **35**, 311-328.

Lindsay, B. (1982). Conditional score functions: Some optimality results. *Biometrika* **69**, 503-12.

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.

Mukerjee, R. and Reid, N. (1999). On confidence intervals associated with the usual and adjusted likelihoods. *J. Roy. Statist. Soc. Ser. B* **61**, 945-953.

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika* **90**, 533-549.

Severini, T. A. (1998). Likelihood functions for the elimination of nuisance parameters. *Biometrika* **85**, 507-522.

Severini, T. A. (2000). *Likelihood Methods in Statistics.* Oxford University Press, Oxford.

Severini, T. A. (2007). Integrated likelihood functions for non-Bayesian inference. *Biometrika* **94**, 529-42.

Sweeting, T. J. (1987). Discussion of the paper by Professors Cox and Reid. *J. Roy. Statist. Soc. Ser. B* **49**, 20-22.

Department of Statistics, Northwestern University, Evanston, IL 60208-4070, U.S.A.

E-mail: severini@northwestern.edu

(Received February 2009; accepted September 2009)