Statistica Sinica **9**(1999), 713-724

ON THE RELATIONSHIP BETWEEN BAYESIAN AND NON-BAYESIAN ELIMINATION OF NUISANCE PARAMETERS

Thomas A. Severini

Northwestern University

Abstract: Consider a statistical model parameterized by a scalar parameter of interest θ and a nuisance parameter λ. Many methods of inference are based on a "pseudo-likelihood" function, a function of the data and θ that has properties similar to those of a likelihood function. Commonly used pseudo-likelihood functions include conditional likelihood functions, marginal likelihood functions, and profile likelihood functions. From the Bayesian point of view, elimination of λ is easily achieved by integrating the likelihood function with respect to a conditional prior density $\pi(\lambda|\theta)$; this approach has some well-known optimality properties. In this paper, we study how close certain pseudo-likelihood functions are to being of Bayesian form. It is shown that many commonly used non-Bayesian methods of eliminating λ correspond to Bayesian elimination of λ to a high degree of approximation.

Key words and phrases: Conditional likelihood, integrated likelihood, marginal likelihood, profile likelihood.

1. Introduction

Consider a statistical model parameterized by a scalar parameter of interest θ and a nuisance parameter λ. The presence of the nuisance parameter λ often makes likelihood inference about θ difficult, from both the practical and theoretical points of view. Hence, many methods have been proposed for eliminating λ from the likelihood function. These methods often lead to the construction of a "pseudo-likelihood", a function of the data and θ with properties similar to those of a likelihood function. Commonly used pseudo-likelihood functions include conditional, marginal, and profile likelihood functions. See, for example, Barndorff-Nielsen and Cox (1994) for a detailed discussion of likelihood-based inference.

From the Bayesian point of view, elimination of λ is easily achieved by "integrating out" λ using $\pi(\lambda|\theta)$, a conditional prior density for λ given θ. The integrated likelihood function is treated as a genuine likelihood function for θ when forming the posterior distribution of θ. The Bayesian method of eliminating λ, along with Bayesian methods in general, has some well-known optimality

properties. The drawback of Bayesian methods, of course, is that the prior distribution must be specified. Hence, non-Bayesian methods of eliminating nuisance parameters are still routinely used. See Basu (1977), Dawid (1980) and Bernardo and Smith (1994, Appendix B) for further discussion.

Any method of eliminating λ that corresponds to integration with respect to a prior density will enjoy the same optimality properties as a genuine Bayesian analysis. Let $\bar{\ell}(\theta)$ denote a pseudo-likelihood function for θ based on the observation of a random variable Y. Then $\bar{\ell}(\theta)$ corresponds to Bayesian elimination of λ provided there exists $\pi(\lambda|\theta)$ such that

$$\bar{\ell}(\theta) = c(y) \int_{\Lambda} \ell(\theta, \lambda)\pi(\lambda|\theta)d\lambda; \tag{1}$$

here c depends only on Y and Λ is the space of possible λ. When (1) holds, $\bar{\ell}(\theta)$ will be said to be (exactly) Bayesian with respect to the prior $\pi(\lambda|\theta)$. As in Berger, Liseo and Wolpert (1997), we will not require $\pi(\lambda|\theta)$ to be a proper density.

Many commonly used pseudo-likelihood functions do not satisfy (1) for any choice of prior density. The goal of this paper is to study how close (1) is to being satisfied for several non-Bayesian methods of eliminating nuisance parameters. One result of this analysis is that it is possible to identify those non-Bayesian methods that are closest to being Bayesian; these methods may be expected to yield good results for a wide variety of models.

This analysis is related to, but different than, the frequency properties of quantiles of the posterior distribution, a problem which has received considerable attention; see, for example, Datta (1996), DiCiccio and Martin (1991, 1993), Datta and Ghosh (1995), Ghosh and Mukerjee (1993), Nicolau (1993), Severini (1991, 1993), Stein (1985) and Tibshirani (1989). Other related results are given by Liseo (1993) and Berger, Liseo and Wolpert (1997) on integrated likelihood functions, Pierce and Peters (1994), Severini (1994), and Berger, Boukai and Wang (1997) on the relationship between Bayesian and non-Bayesian methods, and Reid (1995) on the relationship between Bayesian and non-Bayesian large-sample theory.

The outline of the paper is as follows. In Section 2, the details of approximately Bayesian elimination of nuisance parameters are presented. These results are applied to conditional and marginal likelihood functions in Section 3 and to profile likelihood functions in Section 4. A brief discussion is given in Section 5.

2. Approximately Bayesian Elimination of Nuisance Parameters

Let $\ell(\theta, \lambda)$ denote the likelihood function for (θ, λ) based on $Y_1, \ldots, Y_n$ and let $\bar{\ell}(\theta)$ denote a pseudo-likelihood function for θ. Given a conditional prior

density $\pi(\lambda|\theta)$, let

$$H(\theta) = \log \int \ell(\theta, \lambda)\pi(\lambda|\theta)d\lambda.$$

We will say that $\bar{\ell}(\theta)$, or $\bar{L} = \log \bar{\ell}(\theta)$, is jth-order asymptotically Bayesian with respect to $\pi(\lambda|\theta)$ if

$$\bar{L}(\theta_0 + \delta/\sqrt{n}) - \bar{L}(\theta_0) = H(\theta_0 + \delta/\sqrt{n}) - H(\theta_0) + O_p(n^{-j/2})$$

for all θ_0; here (θ_0, λ_0) denotes the true parameter point. In this case, the pseudo-likelihood function $\bar{\ell}(\theta)$ is locally equivalent to an integrated likelihood function to order $O_p(n^{-j/2})$.

We assume that the sequence of log-likelihood functions $\{L(\theta, \lambda)\}$ is Laplace-regular in the sense of Kass, Tierney and Kadane (1990) and all prior densities are assumed to be six-times differentiable. In addition, assume that $(\hat{\theta}, \hat{\lambda})$, the maximum likelihood estimator of (θ, λ), is consistent and asymptotically normally distributed with covariance matrix $I(\theta, \lambda)^{-1}$ where $I(\theta, \lambda)$ denotes the Fisher information matrix which is positive-definite and differentiable. These same conditions are assumed to hold for $\hat{\lambda}_\theta$, the maximum likelihood estimator of λ for fixed θ. Sufficient conditions for these results are given, e.g., by Lehmann ((1983), Section 6.4) in the i.i.d. case and by Amemiya ((1985), Section 4.1) in more general settings. Finally, all pseudo-likelihood functions are assumed to be three-times differentiable functions of θ.

Under these regularity conditions,

$$H(\theta) \equiv \log \int \ell(\theta, \lambda)\pi(\lambda|\theta)d\lambda = c(Y) + \frac{1}{2}\hat{\psi}(\theta) + \log \pi(\hat{\lambda}_\theta|\theta) + L_p(\theta) + R_n(\theta) \quad (2)$$

where $c(Y)$ depends only on Y, $\hat{\psi}(\theta) = \log|\Sigma_\theta|$ and $-\Sigma_\theta$ is the inverse of the Hessian of $L(\theta, \lambda)$ with respect to λ evaluated at $\hat{\lambda}_\theta$ (Kass, Tierney and Kadane (1990)). Here $R_n(\theta)$ satisfies $R_n(\theta) = O_p(n^{-1})$ and $R_n(\theta_0 + \delta/\sqrt{n}) - R_n(\theta_0) = O_p(n^{-3/2})$.

3. Conditional and Marginal Likelihood Functions

In some cases, a pseudo-likelihood function may be based on either the marginal or conditional distribution of a given statistic. For instance, suppose that the minimal sufficient statistic model may be written $S = (T, A)$ such that the density of S satisfies $p(s; \theta, \lambda) = p(t|a; \theta)p(a; \theta, \lambda)$; throughout the paper density functions will be denoted by the generic symbol p with the argument indicating the density function under consideration. In this case, a pseudo-likelihood function may be based on $p(t|a; \theta)$. The function $p(t|a; \theta)$ is Bayesian with respect to $\pi(\lambda|\theta)$ provided

$$\int p(a; \theta, \lambda)\pi(\lambda|\theta)d\lambda$$

does not depend on θ for each a. This case was studied by Severini (1995) who called such a statistic a Bayes-ancillary statistic for θ in the presence of λ.

This condition is satisfied if A is S-ancillary. A statistic A is said to be S-ancillary for θ in the presence of λ if the family of density functions $\{p(a;\theta,\lambda) : \lambda \in \Lambda\}$ is the same for each θ (e.g., Barndorff-Nielsen (1978)). The following result is given in Dawid (1980).

Proposition 3.1. *If S, the minimal sufficient statistic of the model, may be written $S = (T, A)$ such that the density of S satisfies $p(s;\theta,\lambda) = p(t|a;\theta)p(a;\theta,\lambda)$ and A is S-ancillary, then the conditional likelihood function given by $p(t|a;\theta)$ is exactly Bayesian.*

Proof. Since A is S-ancillary there exists a function $h(\lambda;\theta_0,\theta_1)$, taking values in Λ, such that for any $\theta_0, \theta_1, \lambda$, $p(a;\theta_0,\lambda) = p\{a;\theta_1, h(\lambda;\theta_0,\theta_1)\}$ for almost all a. Fix θ_1; the distribution of A depends only on the parameter $\phi = h(\lambda;\theta,\theta_1)$ which takes values in the set Λ. Hence, A is Bayes-ancillary with respect to any prior under which θ and ϕ are independent.

One situation in which conditional inference is often used is in inference about the canonical parameter of an exponential family model. Suppose that T is a scalar statistic and A is a vector statistic with the same dimension as λ such that the density of (T, A) is of the form

$$\exp\{n\theta t + n\lambda^T a - nd(\theta,\lambda) + Q(t,a)\}. \tag{3}$$

Let $d_{\lambda\lambda}(\theta,\lambda) = \partial^2 d(\theta,\lambda)/\partial\lambda^2$.

Proposition 3.2. *Suppose the density of the minimal sufficient statistic $S = (T, A)$ is of the form (3). Then the conditional likelihood function based on the conditional density of T given A is third-order Bayesian with respect to the prior density $\pi(\lambda|\theta) = |d_{\lambda\lambda}(\theta,\lambda)|$.*

Proof. By Pace and Salvan ((1992), Section 5) the conditional likelihood function satisfies

$$\bar{L}(\theta) = n\theta T + n\hat{\lambda}_\theta^T A - nd(\theta,\hat{\lambda}_\theta) + \frac{1}{2}\log|d_{\lambda\lambda}(\theta,\hat{\lambda}_\theta)| + B(Y) + O_p(n^{-3/2}),$$

where $B(Y)$ depends only on the data. Let $H(\theta)$ denote the integrated likelihood function with respect to a prior $\pi(\lambda|\theta)$. Since $L_p(\theta) = n\theta T + n\hat{\lambda}_\theta^T A - nd(\theta,\hat{\lambda}_\theta)$,

$$\bar{L}(\theta) = H(\theta) + \log|d_{\lambda\lambda}(\theta,\hat{\lambda}_\theta)| - \log\pi(\hat{\lambda}_\theta|\theta) + B_1(Y) + O_p(n^{-3/2}),$$

where $B_1(Y)$ depends only on Y, yielding the result.

Note that $d_{\lambda\lambda}(\theta,\lambda)$ is $I_{\lambda\lambda}(\theta,\lambda)$, the Fisher information matrix for λ for fixed θ. Hence, the effective prior density is the square of the conditional prior density

of λ given θ that is often recommended, $|I_{\lambda\lambda}(\theta,\lambda)|^{1/2}$. Although this result is somewhat surprising, there is another interpretation of $|d_{\lambda\lambda}(\theta,\lambda)|$. In the model (3), consider the nuisance parameter $\phi = E(A;\theta,\lambda)$; the maximum likelihood estimate of ϕ for fixed θ is independent of θ, so that θ and ϕ are unrelated in a certain sense. The prior $|d_{\lambda\lambda}(\theta,\lambda)|$ corresponds to a uniform prior on ϕ.

Example 1. *Poisson regression*

Let $Y_1,\ldots,Y_n$ denote independent Poisson random variables such that Y_j has mean $\exp\{\lambda+\theta X_j\}$, where $X_1,\ldots,X_n$ are known constants. The conditional distribution of the data given $A=\sum Y_j$ does not depend on λ and the conditional likelihood function is given by

$$\ell_c(\theta) = \frac{\exp\{\theta\sum X_jY_j\}}{[\sum\exp\{\theta X_j\}]^A}.$$

Since A is Poisson with mean $\phi \equiv \sum\exp\{\lambda+\theta X_j\}$, A is S-ancillary and, by Proposition 3.1, $\ell_c(\theta)$ is exactly Bayesian with respect to any prior such that ϕ and θ are independent.

Example 2. *Comparison of binomial probabilities*

Let Y and X denote independent binomial random variables each with index n and success probabilities p and q, respectively. Let $\theta=\log\{p(1-q)/[q(1-p)]\}$ and take $\lambda=\log(\frac{q}{1-q})$. The conditional distribution of Y,X given $A=X+Y$ does not depend on λ and the conditional likelihood function is given by

$$\ell_c(\theta) = \frac{\exp\{\theta Y\}}{\sum_{j=0}^{\min(n,A)}\binom{n}{j}\binom{n}{a-j}\exp\{j\theta\}}.$$

According to Proposition 3.2, $\ell_c(\theta)$ is third-order asymptotically Bayesian with respect to

$$\pi(\lambda|\theta) = \frac{\exp\{\theta+\lambda\}}{(1+\exp\{\theta+\lambda\})^2} + \frac{\exp\{\lambda\}}{(1+\exp\{\lambda\})^2}.$$

Another approach to eliminating a nuisance parameter is to use a marginal likelihood function. Suppose that the distribution of a statistic T does not depend on λ; then inference about θ may be based on the marginal distribution of T. The function $p(t;\theta)$ is Bayes with respect to $\pi(\lambda|\theta)$ provided the following integral does not depend on θ:

$$\int p(a|t;\theta,\lambda)\pi(\lambda|\theta)d\lambda.$$

One common class of models for which a marginal likelihood function exists is composite transformation models. Here we give only a brief overview of transformation models; for more details see Barndorff-Nielsen, Blæsild and Eriksen

(1989). In particular, we assume that the model is a standard composite transformation model in the sense of Barndorff-Nielsen, Blæsild and Eriksen ((1989), Section 8). Let X denote a random variable taking values in $\mathcal{X}$ and let $\mathcal{G}$ denote a group of transformations $g : \mathcal{X} \mapsto \mathcal{X}$ such that each $g \in \mathcal{G}$ is one-to-one and onto. Let $\mathcal{P} = \{P_\eta : \eta \in H\}$ denote a set of probability measures containing the distribution of X. We will say that $\mathcal{P}$ is a transformation model with group $\mathcal{G}$ if for each $\eta \in H$, $g \in \mathcal{G}$, there exists a unique $\eta^* \in H$ such that $P_\eta(B) = P_{\eta^*}(gB)$ for all measurable sets B. We denote η^* by $g\eta$ and, hence, we consider $\mathcal{G}$ as a group of transformations acting on H as well.

Let $\mathcal{P} = \{P_{\theta,\lambda} : \theta \in \Theta, \lambda \in \Lambda\}$ denote a set of probability distributions of a random variable Y. This model is called a composite transformation model with index parameter θ and group parameter λ if $\mathcal{P}$ may be written as $\mathcal{P} = \{\mathcal{P}_\theta : \theta \in \Theta\}$ such that each $\mathcal{P}_\theta = \{P_{\theta,\lambda} : \lambda \in \Lambda\}$ forms a transformation model with group $\mathcal{G}$. The following result is discussed in Barndorff-Nielsen, Blæsild and Eriksen ((1989), Section 8).

Proposition 3.3. *Suppose that the family of distributions of Y forms a composite transformation model with index parameter θ and group parameter λ. Then the marginal likelihood function based on the maximal invariant statistic is exactly Bayesian with respect to the right-invariant measure on $\mathcal{G}$, the group of transformations.*

Example 3. *Exponential regression*

Let $Y_1, \ldots, Y_n$ denote independent exponential random variables such that Y_j has mean $[\lambda \exp\{\theta X_j\}]^{-1}$, where $X_1, \ldots, X_n$ are fixed constants. The marginal likelihood based on the distribution of the maximal invariant statistic $(Y_2/Y_1, \ldots, Y_n/Y_1)$ is given by

$$\frac{\exp\{\theta \sum X_j\}}{[\sum Y_j \exp\{\theta X_j\}]^n}. \tag{4}$$

This model is clearly a composite transformation model with λ as the group parameter; the right-invariant measure has density λ^{-1}. It is straightforward to show that the integrated likelihood function with respect to this density is exactly the marginal density (4). Hence, according to Proposition 3.3, the marginal likelihood function is exactly Bayesian.

Example 4. *Dispersion parameter in a normal-theory linear model*

Let Y denote an n-dimensional multivariate normal random variable with mean vector $X\beta$ and covariance matrix $\Sigma(\theta)$; here X is an $n \times p$ matrix of full rank. This is a composite transformation model with group parameter β. The maximal invariant statistic is $A = Y - P_X Y$ where P_X is a linear transformation

representing projection onto the linear space spanned by the columns of X. The marginal likelihood function based on A is given by

$$\frac{\exp\{-\frac{1}{2}A'\Sigma(\theta)^{-1}A+\frac{1}{2}A\Sigma(\theta)^{-1}X(X'\Sigma(\theta)^{-1}X)^{-1}X'\Sigma(\theta)^{-1}A\}}{|\Sigma(\theta)|^{1/2}|X'\Sigma(\theta)^{-1}X|^{1/2}}$$

(e.g., McCullagh and Nelder (1989), Section 7.2). According to Proposition 3.3, this marginal likelihood function is exactly Bayesian with respect to the right-invariant prior density 1.

4. Profile Likelihood and Related Methods

The pseudo-likelihood functions considered in the previous section can only be applied in special circumstances. In this section, we analyze some pseudo-likelihood functions that are generally applicable. We first consider the properties of the profile log-likelihood function $L_p(\theta)=L(\theta,\hat{\lambda}_\theta)$ where $\hat{\lambda}_\theta$ denotes the maximum likelihood estimate of λ for fixed θ. Let

$$\mu_{\lambda\lambda\theta}(\theta,\lambda)=\operatorname{plim}\frac{1}{n}E_{\theta,\lambda}[\frac{\partial^3 L}{\partial\lambda^2\partial\theta}(\theta,\lambda)].$$

Proposition 4.1. *The profile likelihood function is second-order asymptotically Bayes with respect to the prior density given by*

$$\pi(\lambda|\theta)=\exp\{-\frac{1}{2}\int^{\theta} tr[I_{\lambda\lambda}(t,\lambda)^{-1}\mu_{\lambda\lambda\theta}(t,\lambda)]dt\}f(\lambda),$$

where θ and λ are orthogonal parameters and $f(\cdot)$ is an arbitrary function of λ.

Proof. Without loss of generality we assume that θ and λ are orthogonal. Using (2), $L_p(\theta)$ is second-order asymptotically Bayesian provided that there exists a prior $\pi(\lambda|\theta)$ such that

$$\frac{d}{d\theta}\log\pi(\hat{\lambda}_\theta|\theta)\mid_{\theta=\theta_0}+\frac{1}{2}\hat{\psi}'(\theta_0)=O_p(n^{-1/2}).$$

Since $\hat{\psi}(\theta)=-\log|-L_{\lambda\lambda}(\theta,\hat{\lambda}_\theta)|$, it follows that

$$\hat{\psi}'(\theta_0)=\operatorname{tr}[I_{\lambda\lambda}(\theta_0,\lambda_0)^{-1}\mu_{\lambda\lambda\theta}(\theta_0,\lambda_0)]+O_p(n^{-1/2}).$$

Let $f(\cdot)$ be arbitrary. Under orthogonality of θ and λ,

$$\frac{d}{d\theta}\log\pi(\hat{\lambda}_\theta|\theta)\mid_{\theta=\theta_0}=\frac{\partial}{\partial\theta}\log\pi(\lambda|\theta)\mid_{(\theta,\lambda)=(\theta_0,\lambda_0)}+O_p(n^{-1/2})$$

so that $L_p(\theta)$ is second-order asymptotically Bayesian by taking $\pi(\lambda|\theta)$ to satisfy

$$\log\pi(\lambda|\theta)=-\frac{1}{2}\int^{\theta}\operatorname{tr}[I_{\lambda\lambda}(t,\lambda)^{-1}\mu_{\lambda\lambda\theta}(t,\lambda)]dt+\log f(\lambda).$$

Although the profile likelihood function is often useful, it is not a genuine likelihood function. For instance, since it treats the nuisance parameter as fixed at the value $\hat{\lambda}_\theta$, it may overstate the amount of information available about θ. Hence, adjusted versions of the profile likelihood are sometimes used; see, for example, Cox and Reid (1987), Barndorff-Nielsen and Cox (1994) and DiCiccio Martin, Stern and Young (1996) for further discussion. For instance, the adjusted profile log-likelihood function proposed by Cox and Reid (1987) is given by $L_{ap}(\theta) = L_p(\theta) + \hat{\psi}(\theta)/2$ where θ and λ are orthogonal parameters. The following property of $\ell_{ap}(\theta)$ was recognized by Sweeting (1987); the proof follows immediately from the expansion (2).

Proposition 4.2. *The adjusted profile likelihood function $\ell_{ap}(\theta)$ is third-order asymptotically Bayes with effective prior density $\pi(\lambda|\theta) = 1$.*

Note that $L_{ap}(\theta)$ has an important drawback: it depends on the orthogonal parameterization used, although any two versions of $L_{ap}(\theta)$ agree to order $O_p(n^{-1})$ for θ of the form $\theta = \theta_0 + O(n^{-1/2})$. Hence, two different versions of $L_{ap}(\theta)$ are both third-order asymptotically Bayesian, although with respect to different prior distributions. Since the effective prior density of λ given θ is uniform, this suggests that in computing $L_{ap}(\theta)$ the model should be parameterized so that it is reasonable to take λ as uniformly distributed.

The adjusted profile likelihood is closely related to the modified profile likelihood, proposed by Barndorff-Nielsen (1983). The modified profile likelihood function does not require an orthogonal parameterization and it is invariant under interest-respecting parameterizations; see, e.g., Barndorff-Nielsen and Cox (1994). The profile log-likelihood function and the modified profile log-likelihood function are locally equivalent to second-order and, hence, the modified profile log-likelihood function is second-order asymptotically Bayesian (Severini (1998b)).

Example 5. *Variance of a normal distribution*

Let $Y_1, \ldots, Y_n$ denote independent normally distributed random variables each with mean λ and variance θ. The profile log-likelihood function for θ is given by

$$L_p(\theta) = -\frac{n}{2}\log\theta - \frac{1}{2\theta}\sum(Y_j - \bar{Y})^2$$

while the adjusted profile log-likelihood is given by

$$L_{ap}(\theta) = -\frac{n-1}{2}\log\theta - \frac{1}{2\theta}\sum(Y_j - \bar{Y})^2.$$

According to Proposition 4.2, $L_{ap}(\theta)$ is third-order asymptotically Bayesian with respect to the prior density $\pi(\lambda|\theta) = 1$; since $L_{ap}(\theta)$ is identical to the marginal

log-likelihood function based on $\sum(Y_j - \bar{Y})^2$, $L_{ap}(\theta)$ is exactly Bayesian with respect to this prior density.

According to Proposition 4.1, the profile log-likelihood function is second-order approximately Bayesian with respect to a prior density of the form

$$\pi(\lambda|\theta) = \exp\{-\frac{1}{2}\int^{\theta} \mathrm{tr}[I_{\lambda\lambda}(t,\lambda_0)^{-1}\mu_{\lambda\lambda\theta}(t,\lambda_0)]dt\}f(\lambda) = f(\lambda)/\sqrt{\theta}.$$

The profile likelihood function is exactly Bayesian with respect to the prior density $\theta^{-1/2}$ corresponding to $f(\lambda) = 1$.

Example 6. *Index of a negative binomial distribution*

Let $Y_1, \ldots, Y_n$ denote independent observations each with density function

$$\frac{\Gamma(\theta + y)}{\Gamma(y+1)\Gamma(\theta)}\frac{\lambda^y\theta^\theta}{(\lambda+\theta)^{y+\theta}}, \qquad y = 0, 1, \ldots,$$

where $\theta > 0$ and $\lambda > 0$. This is a negative binomial distribution with mean λ and variance $\theta + \theta^2/\lambda$; the parameters λ and θ are orthogonal.

The profile log-likelihood function for θ is given by

$$L_p(\theta) = \sum \log\Gamma(\theta + Y_j) - n\log(\theta) + n\theta\log(\theta) - n(\bar{Y} + \theta)\log(\bar{Y} + \theta).$$

The adjusted profile log-likelihood function is given by

$$L_{ap}(\theta) = \sum \log\Gamma(\theta+Y_j) - n\log(\theta) + (n\theta - 1/2)\log(\theta) - (n\bar{Y} + n\theta - 1/2)\log(\bar{Y}+\theta).$$

It is straightforward to show that the conditional distribution of the data given $\bar{Y}$ depends only on θ; the conditional log-likelihood function given $\bar{Y}$ is given by

$$L_c(\theta) = \sum \log\Gamma(\theta + Y_j) - n\log(\theta) + \log\Gamma(n\theta) - \log\Gamma(n\theta + n\bar{Y}).$$

According to Proposition 4.1, $L_p(\theta)$ is second-order asymptotically Bayes with respect to a prior density of the form

$$\pi(\lambda|\theta) = [\exp\{-\frac{1}{2}\int^{\theta} \mathrm{tr}[I_{\lambda\lambda}(t,\lambda_0)^{-1}\mu_{\lambda\lambda\theta}(t,\lambda_0)]dt\}]^{\frac{1}{2}}f(\lambda) = \sqrt{\frac{\theta}{\theta+\lambda}}f(\lambda).$$

For instance, if we take $f(\lambda) = \lambda^{-1/2}$ we have that $L_p(\theta)$ is second-order asymptotically Bayes with respect to the prior density

$$\pi_1(\theta|\lambda) = [\frac{\theta}{\lambda(\theta+\lambda)}]^{\frac{1}{2}}.$$

This is the Jeffreys' prior density for λ for fixed θ. Direct computation shows that the profile likelihood function is, in fact, third-order asymptotically Bayesian with respect to π_1.

According to Proposition 4.2, $L_{ap}(\theta)$ is third-order asymptotically Bayesian with respect to the prior density $\pi_2(\lambda|\theta) = 1$. The conditional log-likelihood function is exactly Bayesian with respect to the prior density $\pi_3(\lambda|\theta) = \lambda^{-1}$. Hence, each of $L_p(\theta)$, $L_{ap}(\theta)$, and $L_c(\theta)$ are equivalent to log-integrated likelihood functions to a high-degree of approximation, although with respect to different prior densities.

5. Discussion

In this paper it has been shown that many commonly used non-Bayesian methods of eliminating a nuisance parameter correspond to integration with respect to some prior density to a high degree of approximation. For instance, the marginal likelihood function in a composite transformation model is exactly Bayesian and the conditional likelihood function for a canonical parameter of an exponential family model and the adjusted profile likelihood function are both approximately Bayesian to order $O_p(n^{-3/2})$. From the Bayesian point of view, these results offer an explanation as to why many *ad hoc* non-Bayesian methods of eliminating a nuisance parameter are often successful in practice.

One difficulty in using Bayesian methods in practice is that the prior density $\pi(\lambda|\theta)$ must be specified. Recent work has shown that standard methods of constructing a default prior for the full parameter (θ, λ), such as Jeffreys' method, do not necessarily work well for inference regarding θ with λ taken as a nuisance parameter; see, for example, Berger and Bernardo (1989) and Kass and Wasserman (1996). This fact has led to the development of more sophisticated methods of constructing a default prior, such as the reference prior of Berger and Bernardo (Bernardo (1979), Berger and Bernardo (1992a, b), Bernardo and Smith (1994, Section 5.4), Ghosh and Mukerjee (1992)). Unfortunately, it is often non-trivial to determine these priors.

Another approach to Bayesian inference regarding θ suggested by the results presented here is to use an easily implemented non-Bayesian method of eliminating λ which corresponds to Bayesian elimination of λ to a high degree of approximation. Bayesian inference may then be based on the resulting pseudo-likelihood function.

Acknowledgement

This work was partially supported by NSF grant DMS-9505799.

References

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press, Cambridge.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, Chichester.

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

Barndorff-Nielsen, O. E. (1991). Likelihood theory. In *Statistical Theory and Modelling* (Edited by D. V. Hinkley, N. Reid and E. J. Snell). Chapman and Hall, London.

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

Barndorff-Nielsen, O. E., Blæsild, P. and Eriksen, P. S. (1989). *Decomposition and Invariance of Measures, and Statistical Transformation Models*. Springer-Verlag, Heidelberg.

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

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

Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of normal means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84**, 200-207.

Berger, J. O. and Bernardo, J. M. (1992a). On the development of reference priors. In *Bayesian Statistics* **4** (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 35-60. University Press, Oxford.

Berger, J. O. and Bernardo, J. M. (1992b). Ordered group reference priors with application to the multinomial problem. *Biometrika* **79**, 25-38.

Berger, J. O., Boukai, B. and Wang, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis. *Statist. Science* **12**, 133-148.

Berger, J. O., Liseo, B. and Wolpert, R. (1997). Integrated likelihood functions for eliminating nuisance parameters. ISDS Discussion Paper 97-01, Duke University.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. Ser. B* **41**, 113-147.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Wiley, Chichester.

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

Datta, G. S. (1996). On priors providing frequentist validity of Bayesian inference for multiple parametric functions. *Biometrika* **83**, 287-298.

Datta, G. S. and Ghosh, J. K. (1995). On priors providing frequentist validity for Bayesian inference. *Biometrika* **82**, 37-45.

Dawid, A. P. (1980). A Bayesian look at nuisance parameters. In *Bayesian Statistics* (Edited by J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), 167-203. University Press, Valencia.

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**, 189-204.

DiCiccio, T. J. and Martin, M. A. (1991). Approximations of marginal tail probabilities for a class of smooth functions with applications to Bayesian and conditional inference. *Biometrika* **78**, 891-902.

DiCiccio, T. J. and Martin, M. A. (1993). Simple modifications for signed roots of likelihood ratio statistics. *J. Roy. Statist. Soc. Ser. B* **55**, 305-316.

Ferguson, H., Cox, D. R. and Reid, N. R. (1991). Estimating equations from modified profile likelihood. In *Estimating Functions* (Edited by V. P. Godambe). Clarendon, Oxford.

Ghosh, J. K. and Mukerjee, R. (1992). Non-informative priors. In *Bayesian Statistics* **4** (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 195-210. University Press, Oxford.

Ghosh, J. K. and Mukerjee, R. (1993). On priors that match posterior and frequentist distribution functions. *Canad. J. Statist.* **21**, 89-96.

Kass, R. E., Tierney, L. and Kadane, J. B. (1990). The validity of posterior expansions based on Laplace's method. In *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard* (Edited by S. Geisser, J. S. Hodges, S. J. Press and A. Zellner), 473-488. North-Holland, Amsterdam.

Kass, R. E. and Wasserman, L. (1996). Formal rules for selecting prior distributions: a review and annotated bibiliography. *J. Amer. Statist. Assoc.* To appear.

Lehmann, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.

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

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd Edition. Chapman and Hall, London.

Nicolau, A. (1993). Bayesian intervals with good frequentist behaviour in the presence of nuisance parameters. *J. Roy. Statist. Soc. Ser. B* **55**, 377-390.

Pace, L. and Salvan, A. (1992). A note on conditional cumulants in canonical exponential families. *Scand. J. Statist.* **19**, 185-191.

Pierce, D. A. and Peters, D. (1994). Higher-order asymptotics and the likelihood principle: One-parameter models. *Biometrika* **81**, 1-10.

Reid, N. (1995). Likelihood and Bayesian approximation methods. In *Bayesian Statistics* **5** (Edited by J. M. Bernardo et al.), 351-368. University Press, Oxford.

Severini, T. A. (1991). On the relationship between Bayesian and non-Bayesian interval estimates. *J. Roy. Statist. Soc. Ser. B* **53**, 611-618.

Severini, T. A. (1993). Bayesian interval estimates which are also confidence intervals. *J. Roy. Statist. Soc. Ser. B* **55**, 533-540.

Severini, T. A. (1994). Approximately Bayesian inference. *J. Amer. Statist. Assoc.* **89**, 242-249.

Severini, T. A. (1995). Information and conditional inference. *J. Amer. Statist. Assoc.* **90**, 1341-1346.

Severini, T. A. (1998a). An approximation to the modified profile likelihood function. *Biometrika* **85**, 403-411.

Severini, T. A. (1998b). Likelihood functions for inference in the presence of a nuisance parameter. *Biometrika* **85**, 507-522.

Stein, C. (1985). On the coverage probability of confidence sets based on a prior distribution. In *Sequential Methods in Statistics. Banach Center Publications* **6**, 485-514. PWN-Polish Scientific Publishers, Warsaw.

Sweeting, T. J. (1987). Discussion. *J. Roy. Statist. Soc. Ser. B* **49**, 20-21.

Tibshirani, R. (1989). Non-informative priors for one parameter of many. *Biometrika* **76**, 604-608.

Department of Statistics, Northwestern University, Evanston, IL 60201-4070, U.S.A.

E-mail: severini@nwu.edu

(Received September 1997; accepted July 1998)