Integrated Likelihood Inference in Multinomial Distributions

Thomas A. Severini

Department of Statistics, Northwestern University, 2006 Sheridan Road, Evanston, IL, 60208, USA.

Corresponding author(s). E-mail(s): severini@northwestern.edu;

Abstract

Consider a random vector $(N_1, N_2, \ldots, N_m)$ with a multinomial distribution such that $\mathrm{E}(N_j;\, \theta) = np_j(\theta)$, $j = 1, \ldots, m$, where $p_1, \cdots, p_m$ are known functions of an unknown d-dimensional parameter, satisfying $p_1(\theta) + \cdots + p_m(\theta) = 1$. This paper considers non-Bayesian likelihood inference for a real-valued parameter of interest $\psi = g(\theta)$, for a known function g, using an integrated likelihood function. The integrated likelihood function is constructed using the *zero-score expectation (ZSE)* parameter, proposed by Severini [25]; thus, the integrated likelihood function has a number of important properties, such as approximate score- and information-unbiasedness. The methodology is illustrated on the problem of inference for the entropy of the distribution.

Keywords: Entropy, Integrated likelihood ratio statistic, Maximum integrated likelihood estimator, Shannon index

1 Introduction

Let $(N_1, N_2, \ldots, N_m)$ denote a random vector with a multinomial distribution such that $\mathrm{E}(N_j;\, \theta) = np_j(\theta)$, $j = 1, \ldots, m$, where $\theta \in \Theta \subset \Re^d$ is an unknown parameter, $p_j : \Theta \to [0,1]$, $j = 1, 2, \ldots, m$ are known functions satisfying $p_1(\theta) + \cdots + p_m(\theta) = 1$, $\theta \in \Theta$, and m and n are known positive integers. The purpose of this paper is to consider non-Bayesian likelihood inference for a real-valued parameter of interest $\psi = g(\theta)$, where $g : \Theta \to \Psi$ is a known function with two continuous derivatives.

Let $\ell(\theta) = \sum_{j=1}^m N_j \log p_j(\theta)$, $\theta \in \Theta$, denote the log-likelihood function of the model. Standard likelihood-based methods of inference for ψ are based on the profile

log-likelihood function, defined by

$$\ell_p(\psi) = \sup_{\theta \in \Theta : g(\theta) = \psi} \ell(\theta). \tag{1}$$

For instance, $\widehat{\psi}$, the maximum likelihood estimator of ψ, is the maximizer of $\ell_p(\psi)$ and the signed likelihood ratio statistic for ψ is given by $R_\psi = \text{sgn}(\widehat{\psi} - \psi)\,(2\,(\ell_p(\widehat{\psi}) - \ell(\psi)))^{\frac{1}{2}}$.

An alternative likelihood-based approach is to use an integrated likelihood function, constructed from the likelihood function by averaging it over the values of the parameter that correspond to a given value of the parameter of interest. Thus, an integrated likelihood function for ψ has the form

$$\int_{\Theta_\psi} L(\theta)\pi(\theta;\ \psi)d\theta. \tag{2}$$

Here $\Theta_\psi = \{\theta \in \Theta : g(\theta) = \psi\}$ and $\pi(\cdot;\ \psi)$ is a weight function on Θ_ψ which is chosen by the analyst and which may depend on ψ.

Clearly, statistical methods based on an integrated likelihood functions are related to methods of Bayesian inference [3, 16, 17] however, they also have desirable properties from the non-Bayesian, i.e., frequentist, perspective [15, 25, 26, 27]. In particular, integrated likelihood methods often outperform methods based on the (unadjusted) profile likelihood (1), particularly when the dimension of the nuisance parameter is large[1, 3, 6, 7, 11, 13, 19, 21, 24].

2 Choosing the Weight Function

An important feature of an integrated likelihood function is that it depends on the weight function used in its construction; moreover, the properties of the integrated likelihood depend on weight function used. Selection of the weight function so that the integrated likelihood has desirable frequency properties is studied in detail in Severini [25], leading to a recommended approach for choosing the weight function, a brief summary of which is given here.

First suppose that θ may be written as $\theta = (\psi, \lambda)$, where λ is a nuisance parameter, taking values in a set Λ; for convenience, assume that λ is real-valued, although the following discussion applies more generally. In this case, an integrated likelihood for ψ has the form

$$\int_\Lambda L(\psi, \lambda)\pi(\lambda;\ \psi)d\theta,$$

where $\pi(\cdot;\ \psi)$ is a weight function on Λ.

Severini [25] shows that the resulting integrated likelihood function has desirable properties if it is based on a nuisance parameter chosen to be "unrelated" to ψ in a certain technical sense and a weight function for that nuisance parameter that does not depend on ψ. For instance, using this approach, the resulting integrated likelihood function for ψ is approximately score- and information-unbiased [25]; furthermore,

these properties hold for essentially any choice of the weight function, provided that it does not depend on ψ [25, 11, 24].

Starting with a given parameterization (ψ, λ), a nuisance parameter ϕ that is unrelated to ψ is given by the solution to the equation

$$\mathrm{E}\left(\ell_\lambda(\psi,\lambda);\ \widehat{\psi},\phi\right) \equiv \mathrm{E}\left(\ell_\lambda(\psi,\lambda);\ \psi_0,\lambda_0\right)\Big|_{(\psi_0,\lambda_0)=(\widehat{\psi},\phi)} = 0, \tag{3}$$

where $\ell_\lambda(\psi,\lambda) = \partial\ell(\psi,\lambda)/\partial\lambda$; that is, fixing the value of $(\psi,\lambda,\widehat{\psi})$ and solving (3) for ϕ yields $\phi(\psi,\lambda;\ \widehat{\psi})$. Alternatively, given ϕ, the corresponding value of λ may be found by solving (3) for λ. The nuisance parameter ϕ is known as the *zero score expectation (ZSE)* parameter; note that it depends on the data, through $\widehat{\psi}$.

Equivalently, the value of λ corresponding to a given value of the ZSE parameter ϕ can be defined as

$$\arg\max_{\lambda} \mathrm{E}\left(\ell(\psi,\lambda);\ \widehat{\psi},\phi\right) \tag{4}$$

[24].

The likelihood function for (ψ,ϕ) may be written $L(\psi,\lambda(\psi,\phi))$ and, given a weight function $\pi(\phi)$ for ϕ, the integrated likelihood is given by

$$\bar{L}(\psi) = \int_\Phi L(\psi,\lambda(\psi,\phi))\ \pi(\phi)d\phi,$$

where Φ is the space of possible ϕ. Note that for obtaining the integrated likelihood evaluated at ψ it is important to be able to find the value of λ corresponding to given values of ϕ and ψ (and $\widehat{\psi}$) rather than to find the value of ϕ corresponding to a given value of λ.

In many cases, the parameter of interest of the model is given as a function of the parameter θ, without an explicit nuisance parameter; in particular, this is often the case when using the type of multinomial model considered here. In Severini [28], the definition of the ZSE parameter is extended to the case in which an explicit nuisance parameter is not available.

Note that, in order to calculate the integrated likelihood, we need to be able to find the value of θ corresponding to a given value of the ZSE parameter. Define the set

$$\Omega_{\widehat{\psi}} = \{\omega \in \Theta : g(\omega) = \widehat{\psi}\}; \tag{5}$$

that is, $\Omega_{\widehat{\psi}}$ is the set of all parameter values θ such that $g(\theta)$ is equal to the MLE of ψ. For models without an explicit nuisance parameter, elements of $\Omega_{\widehat{\psi}}$ play the role of $(\widehat{\psi},\phi)$, $\phi \in \Lambda$, in models with an explicit nuisance parameter.

Now suppose that an element ω in $\Omega_{\widehat{\psi}}$ is given. Then, generalizing (4), the corresponding value of θ may be described as the maximizer of $\mathrm{E}\left(\ell(\theta);\ \omega\right)$ subject to the restriction that $g(\theta) = \psi$. This characterization yields a function from $\Omega_{\widehat{\psi}}$ to the set

$$\{\theta \in \Theta : g(\theta) = \psi\}. \tag{6}$$

Thus, ω plays the role of the ZSE parameter which, in this context, is subject to the restriction that $g(\omega) = \widehat{\psi}$. That is, ω corresponds to $(\widehat{\psi}, \phi)$, where ϕ is the ZSE parameter based on a specific parameterization of the model.

Writing $\theta = b(\omega, \psi;\ \widehat{\psi})$ for some function b, it follows that the likelihood function for ϕ is given by $L(b(\omega, \psi;\ \widehat{\psi}))$. The corresponding integrated likelihood function evaluated at ψ is obtained by integrating $L(b(\omega, \psi;\ \widehat{\psi}))$ with respect to ω over the set $\Omega_{\widehat{\psi}}$.

It may be shown that the properties of the resulting integrated likelihood are identical to those of an ZSE-parameter-based integrated likelihod when there is an explicit nuisance parameter. See Severini [28] for further details.

3 Application to Multinomial Models

We now construct an integrated likelihood based on the ZSE parameterization for the multinomial model described in the introduction. Suppose that $(N_1, N_2, \ldots, N_m)$ has a multinomial distribution such that $\mathrm{E}(N_j;\ \theta) = np_j(\theta)$ where θ is an unknown d-dimensional parameter, taking values in a set Θ. Consider inference for a real-valued parameter of interest $\psi = g(\theta)$.

An important special case of this model has $d = m$ and $p_j(\theta) = \theta_j$, $j = 1, \ldots, m$; thus, Θ is the probability simplex in $\Re^m$, the set of $\theta_j > 0$, $j = 1, \ldots, m$, satisfying $\theta_1 + \cdots + \theta_m = 1$.

First note that if we can reparameterize the model in terms of a parameter ψ, λ, then the ZSE parameter can be found using the methods described in Severini [25] and Schumann et al [24] and, given a weight function for the ZSE parameter that does not depend on ψ, the integrated likelihood function for ψ can be calculated using one of the computational methods described in Zhao and Severini [29], such as those given by Chib and Jeliazkov [9] and Chib [8].

Here we consider the case in which such a parameterization is not readily available; thus, our goal is to construct an integrated likelihood function for the function of θ given by ψ, without the use of an explicit nuisance parameter.

Let

$$\Psi = \{t \in \Re : g(\theta) = t \quad \text{for some} \quad \theta \in \Theta\}$$

denote the parameter space of ψ and for $t \in \Psi$ define

$$\Omega_t = \{\omega \in \Theta : g(\omega) = t\}.$$

Thus, Ω_t denotes the subset of Θ consisting of those elements of the parameter space Θ for which the corresponding value of $\psi \equiv g(\theta)$ is t.

The following procedure can be used to construct an integrated likelihood function for ψ based on the ZSE parameterization. Let $(n_1, n_2, \ldots, n_m)$ denote the observed value of $(N_1, N_2, \ldots, N_m)$, let

$$L(\theta) = \prod_{j=1}^{m} p_j(\theta)^{n_j} \quad \text{and} \quad \ell(\theta) = \log L(\theta) \tag{7}$$

denote the likelihood and log-likelihood functions, respectively, corresponding to that observation, and let $\widehat{\psi} = g(\widehat{\theta})$ where $\widehat{\theta}$ is the function of $n_1, \ldots, n_d$ maximizing $\ell(\theta)$ over $\theta \in \Theta$. We will also use $L(\theta)$ and $\ell(\theta)$ to denote the random-variable forms of the likelihood and log-likelihood functions,

$$L(\theta) = \prod_{j=1}^{m} p_j(\theta)^{N_j} \quad \text{and} \quad \ell(\theta) = \log L(\theta); \tag{8}$$

when using the symbol $L(\theta)$ or $\ell(\theta)$ it should always be clear from the context if it refers to (7) or (8).

The integrated likelihood for ψ has the general form

$$\bar{L}(\psi) = \int_{\Omega_{\widehat{\psi}}} L(\theta(\omega;\, \psi))\, \pi(\omega) d\omega, \quad \psi \in \Psi,$$

where $\pi(\cdot)$ is a weight function not depending on ψ and, given ω, $\theta(\omega;\, \psi)$ is value of $\theta \in \Theta$ that maximizes $\mathrm{E}(\log L(\theta);\, \omega)$ subject to the restriction that $g(\theta) = \psi$.

Because, for the multinomial model, $\mathrm{E}(N_j;\, \omega) = n p_j(\omega)$ it follows that

$$\mathrm{E}(\ell(\theta);\, \omega) = n \sum_{j=1}^{d} p_j(\omega) \log p_j(\theta).$$

Therefore, for given values of ω and ψ, $\theta(\omega;\, \psi)$ is the value of θ solving the maximization problem

$$\max \sum_{j=1}^{d} p_j(\omega) \log p_j(\theta) \ \text{ over } \ \theta \in \Theta \ \text{ subject to } \ g(\theta) = \psi. \tag{9}$$

To calculate the integrated likelihood we will use Monte Carlo integration. There are two features of the multinomial model that complicate the integration. One is that space over which the integration takes place is the $d-1$ dimensional manifold in $\Re^d$ given by

$$\Omega_{\widehat{\psi}} = \{\omega \in \Theta : g(\omega) = \widehat{\psi}\}.$$

The other is that, for a given value of $\omega_1 \in \Omega_{\widehat{\psi}}$, evaluation of the integrated likelihood of ω_1 requires solution of the constrained maximization problem (9). Fortunately, both of these issues are easy to deal with using simple Monte Carlo integration.

The first step is to draw random variates from $\Omega_{\widehat{\psi}}$, according to a distribution not depending on ψ. An important aspect of this is that the exact distribution used is largely irrelevant; see Severini [25] and Schumann et al [24] for further discussion.

In the present context, this step can be performed by drawing a vector of random variates $(u_1, u_2, \ldots, u_m)$ from the probability simplex in $\Re^m$, according to some distribution not depending on ψ. The a random variate in $\Omega_{\widehat{\psi}}$ can be obtained as the

value of ω that solves

$$\max \sum_{j=1}^{m} u_j \log p_j(\omega) \text{ over } \omega \in \Theta \text{ subject to } g(\omega) = \widehat{\psi}; \tag{10}$$

clearly, the result is function of $(u_1, u_2, \ldots, u_m)$ and an element of $\Omega_{\widehat{\psi}}$. Furthermore, although the probability distribution of the solution to (10) is difficult (or impossible) to determine, by construction, that distribution does not depend on ψ. Note that it is the fact that the exact form of weight function used to construct the integrated likelihood function is unknown that makes the use of other computation methods, like the ones discussed in Zhao and Severini [29], difficult to implement effectively.

For the distribution of $(u_1, u_2, \ldots, u_m)$, here we use the uniform distribution on the simplex, i.e., the symmetric Dirichlet distribution with the parameter value taken to be 1. Let $\tilde{\omega}$ denote a random variate in $\Omega_{\widehat{\psi}}$ found by this method. Then likelihood function for ω evaluated at $\tilde{\omega}$ is given by

$$L(\theta(\tilde{\omega}, \psi)) = \sum_{j=1}^{m} n_j \log p_j(\theta(\tilde{\omega}, \psi))$$

where $\theta(\tilde{\omega}, \psi)$ is the value of θ that maximizes

$$\mathrm{E}\left(\log L(\theta);\ \tilde{\omega}\right) \quad \text{subject to} \quad g(\theta) = \psi.$$

Consider calculation of $\bar{L}(\psi_1)$ for a specific value $\psi_1 \in \Psi$. For a given value of R, let $\omega_1, \omega_2, \ldots, \omega_R$ denote random variates found by the method described above. Let $L_j = L(\theta(\omega_j;\ \psi_1)),\ j = 1, \ldots, R,$, where $\theta(\omega_j;\ \psi_1)$ is the solution to (9), with $\omega = \omega_1$ and $\psi = \psi_1$. Then $\bar{L}(\psi_1)$ is given by $\sum_{j=1}^{R} L_j / R$, for large R.

4 Inference for the Entropy of a Distribution

We now consider a specific example. Suppose that $(N_1, N_2, \ldots, N_m)$ has a multinomial distribution, as described in Section 1, with $p_j(\theta) = \theta_j$, $j = 1, \ldots, m$ and Θ taken to be the probability simplex in $\Re^m$. Let ψ denote the entropy of the distribution, given by

$$\psi = -\sum_{j=1}^{m} \theta_j \log(\theta_j),$$

with $\log(\cdot)$ indicating the natural logarithm and $0 \log(0)$ taken to be 0. Note, although the minimum value of ψ is always 0, the maximum value is $\log(m)$.

To calculate the value of the integrated likelihood function for a specific value of ψ, ψ_1, using the method described in Section 3, we use the following steps.

1. Draw random variates $(u_1, \ldots, u_m)$ from the m-dimensional probability simplex.

2. Maximize $\sum_{j=1}^{m} u_j \log(\omega_j)$ with respect to $\omega_j > 0$, $j = 1, \ldots, m$ subject to the restrictions that

$$\sum_{j=1}^{m} \omega_j = 1 \quad \text{and} \quad -\sum_{j=1} \omega_j \log(\omega_j) = \widehat{\psi}.$$

   Let $(\widehat{\omega}_1, \ldots, \widehat{\omega}_m)$ denote the solution to this maximization problem; note that this is a random variate drawn from the set $\Omega_{\widehat{\psi}}$, with a distribution not depending on ψ.
3. Find $(\widehat{\theta}_1, \ldots, \widehat{\theta}_m)$ as the maximizer of $\sum_{k=1}^{n} \widehat{\omega}_k \log(\theta_k)$, with respect to $(\theta_1, \ldots, \theta_m)$ subject to the restrictions that $\theta_j \geq 0$, $j = 1, \ldots, m$,

$$\sum_{j=1}^{m} \theta_j = 1, \quad \text{and} \quad -\sum_{j=1} \theta_j \log(\theta_j) = \psi_1.$$

   Note that $(\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ is a function of $(\widehat{\omega}_1, \ldots, \widehat{\omega}_m)$ and ψ_1.
4. Define

$$L_1 = \prod_{k=1}^{m} \widehat{\theta}_k^{n_k};$$

   note that this is a function of ψ_1.
5. Repeat steps (1) – (4) $R - 1$ times to obtain $L_1, \ldots, L_R$. Calculate $\bar{L}(\psi_1)$, the integrated likelihood function corresponding to $\psi = \psi_1$ by

$$\frac{1}{R} \sum_{j=1}^{R} L_j.$$

This procedure can be repeated for different values of ψ_1, as needed. Note that steps (1) and (2) do not use the value of ψ_1, so the results from those steps should be used for multiple values of ψ_1; in addition to being more computationally efficient, reusing the results in this way leads to a smoother form for the integrated likelihood.

A simple way to draw a sample from the m-dimensional probability simplex is to draw random variates from a non-negative distribution and then divide those random variates by their sum; using the standard exponential distribution as the non-negative distribution yields the uniform distribution on the probability simplex [12, Chapter XI].

To carry out steps (2) and (3) a suitable constrained optimization algorithm must be used. The results presented in this paper use the augmented Lagrangian method of Conn et al [10] and Birgin and Martinez [4], as implemented in the NLopt optimization library [14] using the R interface [22].

The profile likelihood has a form that is similar to that of the integrated likelihood. The profile likelihood is given by

$$\prod_{k=1}^{m} \widehat{\theta}_k^{n_k}$$

where $\widehat{\theta}_1, \ldots, \widehat{\theta}_k$ is the solution to the maximization problem given in step (3), except that, for the profile likelihood, $\widehat{\omega}_k$ is replaced by N_k/n, $k = 1, \ldots, m$.

Note that it is well-known that the maximum likelihood estimator of the entropy is negatively biased; e.g., Miller [20] shows that the bias of the maximum likelihood estimator has an expansion of the form

$$-\frac{m-1}{2n} + O(\frac{1}{n^2}) \quad \text{as} \quad n \to \infty. \tag{11}$$

5 Some Examples

5.1 Introduction

In this section, the methodology described in the previous section is applied to three sets species abundance data. In this context, the entropy is a measure of species diversity, known as the Shannon index. See, for example, Magurran [18] for a detailed discussion of biological diversity measures in general and the Shannon index in particular. For each set of data the integrated likelihood function is presented, along with the point and interval estimates of the Shannon index. These results are compared to the corresponding results based on the profile likelihood. In addition, for each set of data, the results of a small simulation study are presented, using random variates drawn from a distribution based on the dataset under consideration. The calculations in this section all use $R = 250$.

5.2 Desert rodents

Brown [5] reports the results of a study of the species diversity of rodents in several western United States locations; here we consider the data from dune 17, located in Utah. Six different species were observed and the (ordered) observed counts are $1, 1, 2, 4, 7, 10$. Thus, for these data $m = 6$ and $n = 25$.

Figure 1 contain plots of the integrated log-likelihood (solid line) and the profile log-likelihood (dashed line) for these data; note that both log-likelihoods have been standardized to have maximum value 0.

For these data, the maximum integrated likelihood estimate of the entropy is 1.544 while the maximum likelihood estimate is 1.476. The integrated likelihood function can also be used to construct an integrated likelihood ratio statistic, which can be used to set confidence limits for the true value of the entropy [26]. Here the approximate 95% integrated likelihood ratio confidence interval is given by $(1.235, 1.765)$; the corresponding likelihood ratio confidence interval is given by $(1.178, 1.705)$.

To investigate the frequency properties of these methods, a small simulation study was conducted. Random variates were drawn from the multinomial distribution with $n = 25$, $m = 6$, and cell probabilities taken to be the empirical relative frequencies based on the rodent data; thus, θ_1 and θ_2 are each taken to be $1/25$, θ_3 is taken to be $2/25$, and so on. It follows that the true value of the entropy for this distribution is the maximum likelihood estimate based on the rodent data, 1.476. The estimated bias, standard deviation, and root mean-squared error (RMSE) based on 1000 Monte

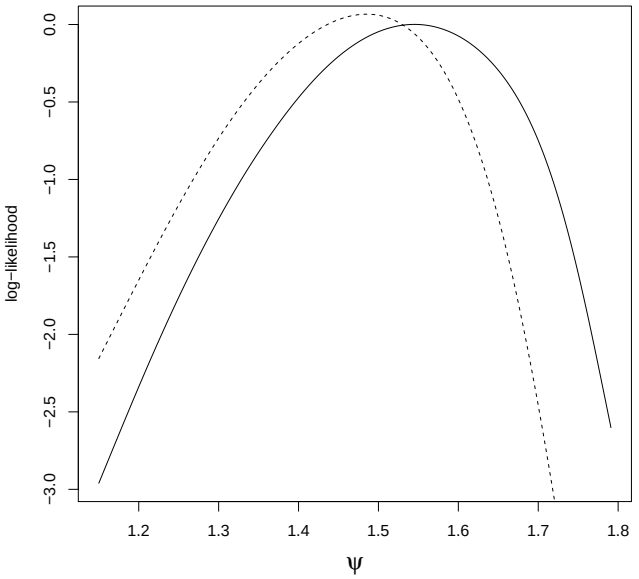

Fig. 1 Integrated (solid) and Profile (dashed) Log-Likelihood Functions for the Rodent Data

Table 1 Simulation Results Based on the Rodent Data

	Likelihood	
	Integrated	Profile
Bias	-0.018	-0.115
SD	0.151	0.156
RMSE	0.153	0.194
Coverage	0.958	0.851
Length	0.547	0.522

Carlo replications are given in Table 1. Thus, the bias of the maximum integrated likelihood estimator appears to be much smaller than that of the maximum likelihood estimator; however, the standard deviations of the two estimators are about equal.

Table 1 also includes the estimated coverage probability and the estimated average length of the approximate 95% integrated-likelihood ratio confidence interval based on the simulation results described previously, along the corresponding values for the approximate 95% likelihood ratio confidence interval. Thus, based on these results, the coverage probability of the integrated-likelihood ratio confidence interval is much closer to the nominal value than is the coverage probability of the likelihood ratio confidence, although it is slightly longer on average. The relatively low coverage probability of the likelihood ratio interval is not surprising, given the large bias of the maximum likelihood estimator.

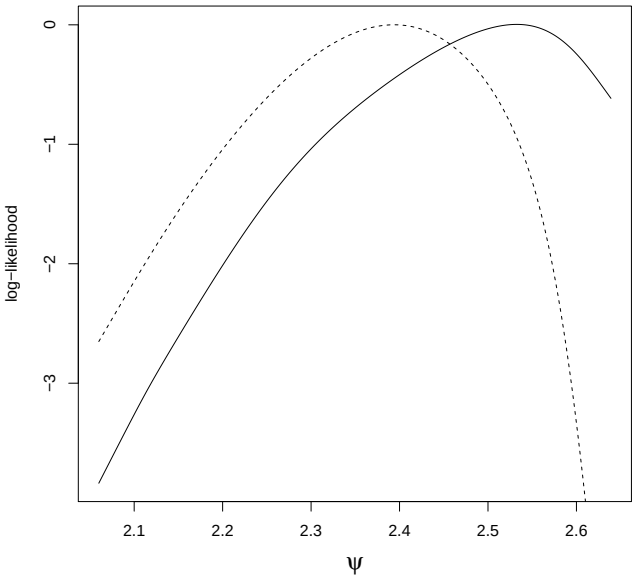

Fig. 2 Integrated (solid) and Profile (dashed) Log-Likelihood Functions for the Balrath Woods Data

5.3 Birds in Balrath Woods

Roycroft et al [23] reports data on the species diversity of birds in 10 transects in Bathrath Woods, Ireland; here we use the data in transect E1. 14 species were observed, with (ordered) observed counts $1,1,1,1,1,2,2,2,2,3,3,3,6,8$. It follows that $m = 14$ and $n = 36$. Thus, compared to the rodent data considered previously, this dataset has more observations but many more cells, leading to many small cell counts; note that the average number of observations per cell is about 2.6 whereas, for the rodent data, the corresponding average is about 4.2.

Figure 2 contains plots of the integrated log-likelihood (solid line) and the profile log-likelihood (dashed line) for these data; both log-likelihoods have been standardized to have maximum value 0. Note that the maximum value of ψ is $\log(14) \doteq 2.639$, which explains the appearance of the integrated log-likelihood plot.

For these data, the maximum integrated-likelihood estimate of the entropy is 2.535 while the maximum likelihood estimate is 2.394. The approximate 95% integrated-likelihood ratio confidence interval is given by $(2.208, 2.639)$; the corresponding likelihood ratio confidence interval is given by $(2.119, 2.572)$. Note that, because $\psi \leq \log(14)$, the integrated likelihood ratio confidence interval is necessarily asymmetric around the maximum integrated likelihood estimate.

To investigate the frequency properties of these methods, a small simulation study was conducted using the same basic approach used in subsection 5.2, but with the distribution used in the simulation based on the Balrath Woods data; thus, here $m = 14$ and $n = 36$ and the true value of the entropy is 2.394.

Table 2 Simulation Results Based on the Balrath Data

	Likelihood	
	Integrated	Profile
Bias	-0.058	-0.197
SD	0.102	0.133
RMSE	0.117	0.238
Coverage	0.930	0.598
Length	0.403	0.472

The simulation results are given in Table 2. For the procedures based on the integrated likelihood, the results are similar in many respects to those based on the rodent data. However, the properties of the maximum likelihood estimator and the likelihood ratio confidence interval are comparatively worse than the properties given in Table 1. For instance, based on the results in Table 2, the bias of the maximum likelihood estimator is greater than its standard deviation and the coverage probability of the approximate 95% likelihood ratio confidence interval is only about 0.6. One reason for this poor performance may be that the number of bins (14) is relatively large compared to the sample size (36). The performance of methods based on the integrated likelihood are also worse for this study than they are in the study described in subsection 5.2, but the large number of bins seems to have less of an effect on the integrated likelihood methods.

5.4 Birds in Killarney woodlands

Batten [2] reports data on the bird species observed in territories in several woodlands in Killarney, Ireland; here we use the data from the Sitka spruce plot. Eight species were observed, with (ordered) observed counts $1, 3, 4, 6, 7, 10, 14, 30$. It follows that, for these data, $m = 8$ and $n = 75$. Thus, compared to the two datasets analyzed previously, this dataset has a larger sample size and larger cell counts, in general.

Figure 3 contains plots of the integrated log-likelihood (solid line) and the profile log-likelihood (dashed line) for these data; both log-likelihoods have been standardized to have maximum value 0.

For these data, the maximum integrated likelihood estimate of the entropy is 1.764 while the maximum likelihood estimate is 1.715. The approximate 95% integrated likelihood ratio confidence interval is given by $(1.564, 1.920)$; the corresponding likelihood ratio confidence interval is given by $(1.520, 1.876)$. Note that, in this example, the results based on the integrated likelihood are very similar to those based on the profile likelihood.

As with the previous two examples, a small simulation study was conducted to investigate the frequency properties of the methods. The distribution used in the simulation is now based on the Killarney data so that, here, $m = 8$, $n = 75$ and the true

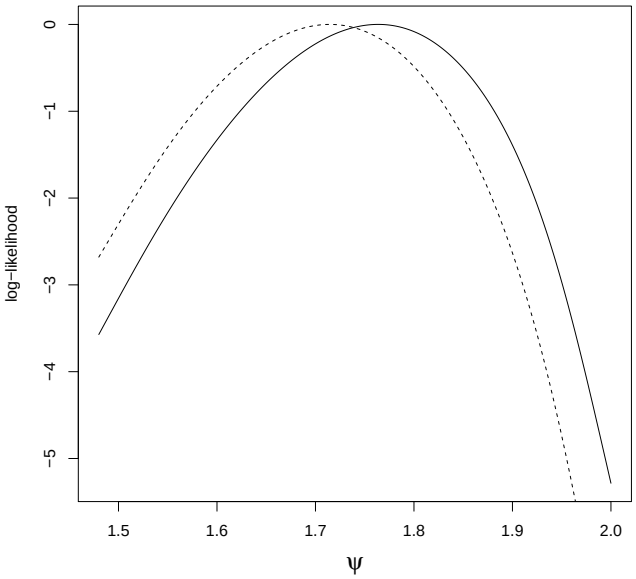

Fig. 3 Integrated (solid) and Profile (dashed) Log-Likelihood Functions for the Killarney Data

Table 3 Simulation Results Based on the Killarney Data

	Likelihood	
	Integrated	Profile
Bias	0.0497	-0.0503
SD	0.0902	0.0927
RMSE	0.103	0.105
Coverage	0.902	0.851
Length	0.342	0.352

value of the entropy is 1.715. Thus, for this study, the sample size is fairly large relative to the number of bins, at least as compared to the studies in subsections 5.2 and 5.3.

The simulation results are given in Table 3. They indicate that, for this distribution of the data, the frequency properties of the procedures based on the integrated and profile likelihoods are in fairly close agreement. This is not surprising because in large samples we expect the integrated and profile likelihoods to be similar [25].

5.5 Summary

The simulation results show that the frequency properties of inferences based on the integrated likelihood are typically better than those of inferences based on the profile likelihood. The advantage of the integrated-likelihood-based methods appears to be

greatest when the sample size is small relative to the number of bins, as illustrated in Table 2. In such cases, the simulation results suggest that integrated-likelihood methods provide an important improvement over methods based on the profile likelihood, such as the maximum likelihood estimator and likelihood ratio confidence intervals.

On the other hand, the results in subsection 5.4 suggest that, when the sample size is large relative to the number of bins, inferences based on the integrated likelihood tend to be close to those based on the profile likelihood. Furthermore, the frequency properties of the two sets of methods are generally similar. Of course, it is important to keep in mind that these conclusions are based on a relatively small simulation study.

Acknowledgments. The comments of the Associate Editor and a referee are gratefully acknowledged.

References

[1] Arellano M, Bonhomme S (2009) Robust priors in nonlinear panel data models. Econometrica 77:489–536

[2] Batten LA (1976) Bird communities for some killarney woodlands. Proceedings of the Royal Irish Academy, Section B: Biological, Geological, and Chemical Science, 76:285–313

[3] Berger JO, Liseo B, Wolpert R (1999) Integrated likelihood functions for eliminating nuisance parameters (with discussion). Statistical Science 14:1–28

[4] Birgin EG, Martinez JM (2008) Improving ultimate convergence of an augmented lagrangian method. Optimization Methods and Software 23:177–195

[5] Brown JH (1973) Species diversity of seed-eating desert rodents in sand dune habitats. Ecology 54:775–787

[6] Carroll RJ, Lombard F (1985) A note on n estimators for the binomial distribution. Journal of the American Statistical Association 80:423–426

[7] Chamberlain G (2007) Decision theory applied to an instrumental variables model. Econometrica 75:609–52

[8] Chib S (1995) Marginal likelihood from gibbs output. Journal of the American Statistical Association 90:1313–1321

[9] Chib S, Jeliazkov I (2001) Marginal likelihood from the metropolis-hastings output. Journal of the American Statistical Association 96:270–2811

[10] Conn AR, Gould NIM, Toint PL (1991) A globally convergent augmented lagrangian algorithm for optimization with general constraints and simple bounds. SIAM J Numer Anal 28:545–572

[11] De Bin R, Sartori N, Severini TA (2015) Integrated likelihoods in models with stratum nuisance parameters. Electronic Journal of Statistics 9:1474–1491

[12] Devroye L (1986) Non-Uniform Random Variate Generation. Springer-Verlag, New York

[13] Ghosh M, Datta GS, Kim D, et al (2006) Likelihood-based inference for ratios of regression coefficients in linear models. Ann Inst Statist Math 58:457–73

[14] Johnson SG (2021) The nlopt nonlinear-optimization package. URL http://github.com/stevengj/nlopt

[15] Kalbfleisch JD, Sprott DA (1970) Application of likelihood methods to models involving large numbers of parameters (with discussion). J R Statist Soc B 32:175–208

[16] Liseo B (1993) Elimination of nuisance parameters with reference priors. Biometrika 80:295–304

[17] Liseo B (2005) The elimination of nuisance parameters. In: Bayesian Thinking: Modeling and Computation, Handbook of Statistics, vol 25. Elsevier/North Holland, Amersterdam, p 193–219

[18] Magurran AE (2004) Measuring Biological Diversity. Blackwell, Oxford

[19] Malley JD, Redner RA, Severini TA, et al (2003) Estimation of linkage and association from allele transformation data. Biometrical Journal 45:349–66

[20] Miller GA (1955) Note on the bias of information estimates. Information Theory in Psychology: Problems and Methods 2:95–100

[21] Osborne JA, Severini TA (2000) Inference for exponential order statistics models based on an integrated likelihood function. Journal of the American Statistical Association 95:1220–1228

[22] R Core Team (2020) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/

[23] Roycroft D, Irwin S, Wilson M, et al (2008) The breeding bird community of balrath wood in 2007. Irish Forestry 65:60–70

[24] Schumann M, Severini TA, Tripathi G (20211) Integrated likelihood based inference for nonlinear panel data models with unobserved effects. Journal of Econometrics 223:73–95

[25] Severini TA (2007) Integrated likelihood functions for non-bayesian inference. Biometrika 94:529–42

[26] Severini TA (2010) Likelihood ratio statistics based on an integrated likelihood. Biometrika 97:481–496

[27] Severini TA (2011) Frequency properties of inferences based on an integrated likelihood function. Statistica Sinica 21(1):433–447

[28] Severini TA (2018) Integrated likelihoods for functions of a parameter. Stat 7:e212

[29] Zhao Z, Severini TA (2017) Integrated likelihood computation methods. Computational Statistics 32:281–313