Electronic Journal of Statistics
Vol. 9 (2015) 1474–1491
ISSN: 1935-7524
DOI: 10.1214/15-EJS1045

Integrated likelihoods in models with stratum nuisance parameters

Riccardo De Bin

Department of Medical Informatics, Biometry and Epidemiology, University of Munich
Marchioninistraße 15, 81377 Munich, Germany
e-mail: debin@ibe.med.uni-muenchen.de

Nicola Sartori

Department of Statistical Sciences, University of Padova
via Cesare Battisti 241, 35121 Padova, Italy
e-mail: sartori@stat.unipd.it

and

Thomas A. Severini

Department of Statistics, Northwestern University
2006 Sheridan Road, 60208 Evanston (IL), USA
e-mail: severini@northwestern.edu

Abstract: Frequentist inference about a parameter of interest in presence of a nuisance parameter can be based on an integrated likelihood function. We analyze the behaviour of inferential quantities based on such a pseudo-likelihood in a two-index-asymptotics framework, in which both sample size and dimension of the nuisance parameter may diverge to infinity. We show that a properly chosen integrated likelihood largely outperforms standard likelihood methods, such as those based on the profile likelihood. These results are confirmed by simulation studies, in which comparisons with modified profile likelihood are also considered.

MSC 2010 subject classifications: Asymptotic properties 62G20.
Keywords and phrases: Modified profile likelihood, non stationary autoregressive model, profile likelihood, profile score bias, two-index asymptotics.

Received September 2014.

Contents

1. Introduction

Consider stratified data $y = (y_1, \ldots, y_q)$, where y_i is a realization of a random variable Y_i of dimension m_i and with density $p_i(y_i; \psi, \lambda_i)$. Suppose that $Y_1, \ldots, Y_q$ are independent and consider ψ as the parameter of interest, with $\lambda = (\lambda_1, \ldots, \lambda_q)$ as a nuisance parameter. We assume that each stratum-specific λ_i has the same meaning and the same parameter space, Λ.

It is well known that in models in which the dimension of the nuisance parameter is large relative to the sample size, many methods of likelihood inference, such as those based on the profile likelihood, can perform poorly. Sometimes, the model structure allows one to base inference on a conditional or marginal likelihood, which is a genuine likelihood for the parameter of interest, not depending on the nuisance parameters, and therefore satisfies all the standard properties of a likelihood (Severini, 2000, Chapter 8). Unfortunately, these special likelihoods are often not available outside special families of distributions and, even when they exist, they may be difficult to compute. To deal with these issues, several modifications to the profile likelihood have been proposed; see Barndorff-Nielsen and Cox (1994, Chapter 8) and Severini (2000, Chapter 9) for general discussion of these methods and further references.

An alternative, and more general, solution is offered by integrated likelihood functions, which are formed by integrating the likelihood function with respect to a weight function for the nuisance parameter (Kalbfleisch and Sprott, 1970); specifically, an integrated likelihood L_I is of the form

$$L_I(\psi) = \int_\Lambda L(\psi, \lambda) g(\lambda; \psi) d\lambda,$$

where $L(\psi, \lambda)$ denotes the likelihood function and $g(\cdot; \psi)$ is a weight function for the nuisance parameter λ. It is not necessary for $g(\cdot; \psi)$ to be a genuine density function. Therefore, in this respect the approach here differs substantially from a random-effects modelization of λ, in which the density for λ_i would typically depend also on additional parameters. Because integrated likelihoods are based on averaging, they often avoid the problems related to maximization that sometimes arise with methods based on the profile likelihood, or its modifications; see, for example, Berger, Liseo and Wolpert (1999). Furthermore, for appropriate choices of the weight function g, integrated likelihoods have a number of attractive frequentist properties (Severini, 2007, 2010, 2011). In addition, the wide availability of reliable routines for numerical integration have made this approach even more appealing.

Inference for ψ based on an integrated likelihood proceeds by treating $L_I(\psi)$ as a genuine likelihood for ψ. In particular, we will study the properties of the usual quantities such as the integrated log likelihood $\ell_I(\psi) = \log L_I(\psi)$, its maximizer $\hat{\psi}_I$, and the integrated likelihood ratio statistic, $W_I = 2\{\ell_I(\hat{\psi}) - \ell_I(\psi)\}$.

In the following, we consider an asymptotic scenario in which both m_i, the within-stratum sample sizes, and q, the number of strata, approaches infinity. This type of "two-index asymptotics" is more relevant to cases in which the number of strata is large relative to the total sample size; see, e.g., Barndorff-Nielsen (1996) for a general discussion of two-index asymptotics and Sartori (2003) for discussion of the properties of profile and modified profile likelihoods in this setting. In this framework, under the frequentist paradigm a thorough analysis of the asymptotic properties of the inferential quantities derived from the integrated likelihood is missing and the aim of this paper is to fill this gap. In practice, our work represents an extension to the results provided by Severini (2007, 2010), who studied integrated likelihoods in the standard one-index asymptotic setting (q fixed).

Although the present paper does not deal with Bayesian inference, we note that the proposed integrated likelihood may be used in conjunction with a prior for ψ in order to obtain an approximate marginal posterior distribution for ψ, as suggested, e.g., in Efron (1993) and Ventura, Cabras and Racugno (2009).

The paper is organized as follows. In Section 2 we discuss the properties of integrated likelihood functions and consider the selection of the weight function. The asymptotic properties of the statistics based on the integrated likelihoods, such as the maximum integrated likelihood estimator and the integrated likelihood ratio statistic, are studied in Section 3. Examples are presented in Section 4, in which comparisons with profile and modified profile likelihoods are considered. Section 5 contains some final remarks.

2. Integrated likelihood functions

Let $L^{(i)}(\psi, \lambda_i) = p_i(y_i; \psi, \lambda_i)$ denote the likelihood function for the ith stratum so that, due to independence between strata, $L(\psi, \lambda) = \prod_{i=1}^{q} L^{(i)}(\psi, \lambda_i)$, with $\lambda = (\lambda_1, \ldots, \lambda_q)$; let $\ell(\psi, \lambda) = \sum_{i=1}^{q} \ell^{(i)}(\psi, \lambda_i)$, where $\ell^{(i)}(\psi, \lambda_i) = \log L^{(i)}(\psi, \lambda_i)$ denotes the log-likelihood function. For the results in the paper it is not necessary for the components of Y_i to be independent. Indeed, some form of dependence among the components of Y_i can also be accommodated, such as in longitudinal data, as long as a central limit theorem for the score statistic associated with $\ell^{(i)}(\psi, \lambda_i)$ can be established; see, for instance, Reid (2003) and Wooldridge (1994). Let $\hat{\lambda}_{i\psi}$ denote the maximum likelihood estimator of λ_i for fixed ψ and let $\ell_P(\psi) = \sum_{i=1}^{q} \ell^{(i)}(\psi, \hat{\lambda}_{i\psi})$ denote the profile log-likelihood. Derivatives will be denoted by subscripts; e.g., $\ell_\psi(\psi, \lambda) = \partial\ell(\psi, \lambda)/\partial\psi$. For notational simplicity, we will consider scalar nuisance parameters; the results are easily extended to the case in which the λ_i are vectors.

In the following, for simplicity of notation in the theoretical development we will assume $m_i = m$, but we could otherwise assume that each m_i can be written

in the form $m_i = K_i\, m$, with $A \leq K_i \leq B$ and where A and B are positive finite numbers (see also Sartori, 2003). This assumption guarantees that the strata sample sizes are asymptotically balanced, in the sense that each m_i is of order $O(m)$, but not $o(m)$. In the examples of Section 4 we will consider the case with m_i that can vary among the strata and in the final section we will discuss further the above assumptions.

A central role in the construction of an integrated likelihood is played by the weight function $g(\lambda_i; \psi)$. We consider weight functions of the form $\prod_{i=1}^{q} g(\lambda_i; \psi)$, leading to an integrated likelihood of the form

$$L_I(\psi) = \prod_{i=1}^{q} \int_{\Lambda} L^{(i)}(\psi, \lambda_i) g(\lambda_i; \psi)\, d\lambda_i.$$

In order to derive the theoretical results when both q and m diverge, we will use the Laplace approximation (see, for instance, Pace and Salvan, 1997, Section 9.3.3), which leads to a manageable asymptotic form for the integral. On the other hand, in practical applications, in particular with moderate values of m, alternative approaches are used to evaluate the integrated likelihood. In particular, even exact integration of the likelihood function may be possible, although this happens only in exceptional cases, as in the examples of Sections 4.1 and 4.3. More generally, we will rely on some form of numerical integration (see, for instance, Press et al., 2007, Chapter 4), as in the binomial example of Section 4.2.

Since each λ_i appears only in a single stratum, an analytic approximation to $\ell_I(\psi)$ can be obtained by using a Laplace approximation in each stratum and then combining the results

$$\ell_I(\psi) = \ell_P(\psi) + \sum_{i=1}^{q} \log g(\hat{\lambda}_{i\psi}; \psi) - \sum_{i=1}^{q} \log\{-\ell_{\lambda_i \lambda_i}(\psi, \hat{\lambda}_{i\psi})\}^{1/2} + O_p(\frac{q}{m});$$

see, e.g., Evans and Swartz (2000, Chapter 4).

The properties of ℓ_I heavily depend on the choice of weight function used in its construction. For instance, in general, the score bias within each stratum is of order $O(1)$ so that, summing across strata,

$$E\{\ell_{I\psi}(\psi); \psi, \lambda\} = O(q).$$

However, suppose that the model is parameterized so that, for each $i = 1, \ldots, q$, ψ and λ_i are orthogonal parameters and the weight function for λ_i does not depend on ψ. Then the score bias within each stratum is of order $O(1/m)$ (Severini, 2007; Sweeting, 1987) so that

$$E\{\ell_{I\psi}(\psi); \psi, \lambda\} = O(\frac{q}{m}).$$

The construction of the orthogonal parametrization can be based on the expected information matrix, as discussed in detail by Cox and Reid (1987), or on the zero-score expectation (ZSE) parametrization, as suggested by Severini

(2007). In the latter case, the solution of the equation

$$E\{\ell_{\lambda_i}(\psi, \lambda_i); \psi_0, \phi_i\} = 0 \tag{2.1}$$

leads to an expression for the ZSE parameter ϕ_i in terms of ψ, λ_i, ψ_0; ψ_0 is then replaced by an estimator, such as the maximum likelihood estimator. Therefore, the ZSE parameter requires a reliable estimator of ψ, which may not be available in the setting considered here. The information-orthogonal parameter requires the solution to a differential equation, which may be difficult to find, and which is not guaranteed to exist unless ψ is a scalar. Hence, in general, the ZSE parameter is easier to obtain; however, in some specific models, the reverse is true. Thus, both approaches are useful in practice and the asymptotic theory presented in this paper applies to either choice.

Hereafter, we refer only to an integrated likelihood function with orthogonal parametrization and weight function not depending on ψ. Although under these conditions the integrated likelihood is relatively insensitive to the specific choice of the weight function, a constant weight function is typically convenient. This integrated likelihood is closely related to the modified profile likelihood (Severini, 2007; Sweeting, 1987). Let $\ell_I^{(i)}(\psi)$ and $\ell_M^{(i)}(\psi)$ denote the integrated and modified profile log-likelihoods, respectively, for the ith stratum. Because of the separability of the nuisance parameters,

$$\ell_I(\psi) = \sum_{i=1}^{q} \ell_I^{(i)}(\psi) \quad \text{and} \quad \ell_M(\psi) = \sum_{i=1}^{q} \ell_M^{(i)}(\psi).$$

Let $\hat{\psi}_M$ denote the maximizer of $\ell_M(\psi)$. The asymptotic properties of $\hat{\psi}_M$ in models with stratum nuisance parameters are considered in Sartori (2003) where it is shown $\hat{\psi}_M = \psi + O_p(1/\sqrt{mq})$ provided that $q/m^3 = o(1)$ and $\hat{\psi}_M = \psi + O_p(1/m^2)$ otherwise. If we define

$$D^{(i)}(\psi) = \ell_I^{(i)}(\psi) - \ell_M^{(i)}(\psi), \quad i = 1, \ldots, q,$$

then

$$\ell_I(\psi) - \ell_I(\hat{\psi}_M) = \ell_M(\psi) - \ell_M(\hat{\psi}_M) + \sum_{i=1}^{q} \{D^{(i)}(\psi) - D^{(i)}(\hat{\psi}_M)\}.$$

By a Taylor's series expansion,

$$\sum_{i=1}^{q} D^{(i)}(\psi) = \sum_{i=1}^{q} D^{(i)}(\hat{\psi}_M) - \sum_{i=1}^{q} D_\psi^{(i)}(\psi)||\hat{\psi}_M - \psi|| + O_p(q||\hat{\psi}_M - \psi||^2), \tag{2.2}$$

where $D_\psi^{(i)}(\psi)$ denotes the first derivative of $D^{(i)}(\psi)$ with respect to ψ. In particular, its sum in q may be written as

$$\sum_{i=1}^{q} D_\psi^{(i)}(\psi) = \sum_{i=1}^{q} E\{D_\psi^{(i)}(\psi); \psi, \lambda_i\} + \sum_{i=1}^{q} [D_\psi^{(i)}(\psi) - E\{D_\psi^{(i)}(\psi); \psi, \lambda_i\}],$$

where the first term on the right-hand side is $O(q/m)$ and the second term is $O_p(\sqrt{q/m})$, due to the fact that $D_\psi^{(i)}(\psi) = O_p(1/\sqrt{m})$ and $E\{D^{(i)}(\psi); \psi, \lambda_i\} = O(1/m)$. These results follow from the relationship between the integrated likelihood and the modified profile likelihood in the single-stratum case.

Substituting in (2.2), we obtain that

$$\sum_{i=1}^{q} D^{(i)}(\psi) = \sum_{i=1}^{q} D^{(i)}(\hat{\psi}_M) + O_p(\frac{q}{m}||\hat{\psi}_M - \psi||) + O_p(\sqrt{\frac{q}{m}}||\hat{\psi}_M - \psi||) + O_p(q||\hat{\psi}_M - \psi||^2),$$

from which, for $\psi = \hat{\psi}_M + O_p(1/\sqrt{mq})$,

$$\ell_I(\psi) = \ell_M(\psi) + O_p(\sqrt{\frac{q}{m^3}}) + O_p(\frac{1}{m}),$$

ignoring terms not depending on ψ. Note that the true value of ψ is of the form $\hat{\psi}_M + O_p(1/\sqrt{mq})$ provided that $q/m^3 = o(1)$.

Similar analyses show that, for $\psi = \hat{\psi}_M + O_p(1/\sqrt{mq})$,

$$\frac{1}{\sqrt{mq}}\ell_{I\psi}(\psi) = \frac{1}{\sqrt{mq}}\ell_{M\psi}(\psi) + O_p(\sqrt{\frac{q}{m^3}}) + O_p(\frac{1}{m}) \tag{2.3}$$

and

$$\frac{1}{mq}\ell_{I\psi\psi}(\psi) = \frac{1}{mq}\ell_{M\psi\psi}(\psi) + O_p(\frac{1}{m}). \tag{2.4}$$

3. Asymptotic properties of statistics based on an integrated likelihood

3.1. Introduction

In this section, the properties of the maximum integrated likelihood estimator and the integrated likelihood ratio statistic are established by showing that these statistics have the same asymptotic distribution as the corresponding statistics based on the modified profile likelihood and then using the results of Sartori (2003), which establishes the properties of inferences based on the modified profile likelihood in models for stratified data.

3.2. Maximum integrated likelihood estimator

Let $\hat{\psi}_I$ denote the maximizer of $\ell_I(\psi)$. Using the usual expansions for the maximizer of a log-likelihood (e.g., Severini, 2000, Ch. 5),

$$\sqrt{mq}(\hat{\psi}_I - \psi) = \frac{\frac{1}{\sqrt{mq}}\ell_{I\psi}(\psi)}{-\frac{1}{mq}\ell_{I\psi\psi}(\psi)} + O_p(\frac{1}{\sqrt{mq}})$$

and

$$\sqrt{mq}(\hat{\psi}_M - \psi) = \frac{\frac{1}{\sqrt{mq}}\ell_{M\psi}(\psi)}{-\frac{1}{mq}\ell_{M\psi\psi}(\psi)} + O_p(\frac{1}{\sqrt{mq}}).$$

It now follows from (2.3) and (2.4) that

$$\sqrt{mq}(\hat{\psi}_I - \psi) = \sqrt{mq}(\hat{\psi}_M - \psi) + O_p(\sqrt{\frac{q}{m^3}}) + O_p(\frac{1}{m}); \tag{3.1}$$

that is, $\hat{\psi}_I$ has the same asymptotic distribution as $\hat{\psi}_M$ provided that $q/m^3 = o(1)$. In particular, under this condition, $\hat{\psi}_I$ has the same asymptotic distribution as the maximum conditional or marginal likelihood estimators, if either exists.

The asymptotic properties of $\hat{\psi}_M$ in models with stratum nuisance parameters are considered in Sartori (2003) and are described in the previous section. Given the relationship between $\hat{\psi}_I$ and $\hat{\psi}_M$, these results may be used to derive the asymptotic properties of $\hat{\psi}_I$. Specifically, $\hat{\psi}_I = \psi + O_p(1/\sqrt{mq})$ provided that $q/m^3 = o(1)$ and $\hat{\psi}_I = \psi + O_p(1/m^2)$ otherwise. Furthermore, when $q/m^3 = o(1)$, $\hat{j}_I^{\frac{1}{2}}(\hat{\psi}_I - \psi)$ is asymptotically normally distributed with error

$$O_p(\sqrt{\frac{q}{m^3}}) + O_p(\frac{1}{m}) + O_p(\frac{1}{\sqrt{mq}});$$

here $\hat{j}_I$ denotes the observed information based on $\ell_I(\psi)$, evaluated at $\hat{\psi}_I$.

For comparison, the maximum likelihood estimator $\hat{\psi}$ satisfies $\hat{\psi} = \psi + O_p(1/\sqrt{mq})$ provided that $q/m = o(1)$ and $\hat{\psi} = \psi + O_p(1/m)$ otherwise. When $q/m = o(1)$, $\hat{j}_P^{\frac{1}{2}}(\hat{\psi} - \psi)$ is asymptotically distributed according to a standard normal distribution with error $O_p(\sqrt{q/m})$; here $\hat{j}_P$ denotes the profile observed information evaluated at $\hat{\psi}$.

In terms of the total sample size $n = mq$ and the number of strata, q, $\hat{j}_I^{\frac{1}{2}}(\hat{\psi}_I - \psi)$ is asymptotically normally distributed provided that $q/n^{\frac{3}{4}} = o(1)$. On the other hand, $\hat{j}_P^{\frac{1}{2}}(\hat{\psi} - \psi)$ is asymptotically normally distributed provided that $q/\sqrt{n} = o(1)$. Furthermore, even when both $\hat{\psi}_I$ and $\hat{\psi}$ are asymptotically normal, the error in the normal approximation to $\hat{\psi}_I$ is smaller than the error in the normal approximation to $\hat{\psi}$. For instance, suppose that $q = n^{\frac{1}{3}}$. Then $\hat{j}_P^{\frac{1}{2}}(\hat{\psi} - \psi)$ is asymptotically normally distributed with error $O_p(1/n^{\frac{1}{6}})$ while $\hat{j}_I^{\frac{1}{2}}(\hat{\psi}_I - \psi)$ is asymptotically normally distributed with error $O_p(1/\sqrt{n})$.

3.3. Integrated likelihood ratio statistics

We now consider likelihood-ratio-type statistics based on an integrated likelihood. Let

$$W_I = 2\{\ell_I(\hat{\psi}_I) - \ell_I(\psi)\}$$

denote the likelihood ratio statistic based on the integrated likelihood. It is straightforward to show that

$$W_I = (\hat{\psi}_I - \psi)^T \hat{j}_I(\hat{\psi}_I - \psi)\{1 + O_p(\hat{\psi}_I - \psi)\}.$$

Using the asymptotic properties of $\hat{j}_I^{\frac{1}{2}}(\hat{\psi}_I - \psi)$ described in the previous subsection, it follows that W_I is asymptotically distributed according to a chi-

squared distribution with p degrees-of-freedom, where p denotes the dimension of ψ, with error

$$O(\sqrt{\frac{q}{m^3}}) + O(\frac{1}{m}) + O(\frac{1}{\sqrt{mq}}),$$

provided that $q/m^3 = o(1)$. That is, W_I is asymptotically distributed according to a chi-squared distribution with p degrees-of-freedom provided that $q/m^3 = o(1)$.

For comparison, the standard likelihood ratio statistic

$$W = 2\{\ell_P(\hat{\psi}) - \ell_P(\psi)\}$$

is asymptotically distributed according to a chi-squared distribution with p degrees-of-freedom provided that $q/m = o(1)$. As is the case for the estimators $\hat{\psi}_I$ and $\hat{\psi}$, even when the condition $q/m = o(1)$ holds, the error in the chi-squared approximation to the distribution W_I is of smaller order than is the error in the chi-squared approximation to the distribution of W.

It is worth noting that, when $p = 1$, similar results hold for the signed likelihood ratio statistic

$$R_I = \mathrm{sgn}(\hat{\psi}_I - \psi)\sqrt{W_I};$$

e.g., when $q/m^3 = o(1)$, R_I is asymptotically distributed according to a standard normal distribution.

4. Examples

4.1. Gamma with common shape parameter

Let Y_{ij}, $i = 1, \ldots, q$, $j = 1, \ldots, m_i$, be independent gamma random variables with shape parameter ψ and scale parameter $1/\lambda_i$, as in Sartori (2003, Example 2). The solution of equation (2.1) gives the ZSE parameter $\phi_i = \hat{\psi}\lambda_i/\psi$, which allows us to derive the integrated log-likelihood with orthogonal parametrization and constant weight function, $\ell_I(\psi) = \sum_{i=1}^{q} \ell_I^{(i)}(\psi)$, where

$$\ell_I^{(i)}(\psi) = \log \int_0^\infty \exp\{\psi \sum_{j=1}^{m_i} \log y_{ij} - \frac{\psi}{\hat{\psi}}\phi_i \sum_{j=1}^{m_i} y_{ij} + m_i \psi \log \frac{\psi}{\hat{\psi}}\phi_i - m_i \log \Gamma(\psi)\}\, d\phi_i.$$

After some algebra, we obtain

$$\ell_I(\psi) = \psi \sum_{i=1}^{q} \{\sum_{j=1}^{m_i} \log y_{ij} - m_i \log \sum_{j=1}^{m_i} y_{ij}\} - \sum_{i=1}^{q} \{m_i \log \Gamma(\psi) - \log \Gamma(m_i \psi)\},$$

which is the conditional log-likelihood for ψ. It is worth noting that the conditional likelihood for the shape parameter of a gamma is also a marginal likelihood. Hence, the same result can be achieved also using as a weight function ϕ_i^{-1}, the weight function related with the right invariant measure (see Pace and Sal-

van, 1997, Example 7.29). Conversely to the above integrated likelihood, in this example the modified profile likelihood is not exactly equivalent to the conditional likelihood (Sartori, 2003, Example 2). The same is true for the integrated likelihood based on the information orthogonal nuisance parameter.

4.2. Matched binomials

Let us consider Y_{i1} and Y_{i2}, $i = 1, \dots, q$, two independent random variables with distribution $Bi(m_i, p_{i1})$ and $Bi(1, p_{i2})$ respectively. Let $\lambda_i = \log\{p_{i1}/(1-p_{i1})\}$ be the stratum nuisance parameter and $\psi = \log\{p_{i2}/(1-p_{i2})\} - \log\{p_{i1}/(1-p_{i1})\}$ be the parameter of interest, common among strata. We may deal with a model like this in case-control studies where we are interested in analyzing the effect of a certain factor by the comparison among one case and m_i controls (Sartori, 2003, Example 3). The likelihood is

$$L(\psi, \lambda) = \prod_{i=1}^{q} \frac{e^{(y_{i1}+y_{i2})\lambda_i + y_{i2}\psi}}{(1+e_i^{\lambda})^{m_i}(1+e^{\psi+\lambda_i})},$$

while the conditional likelihood is a noncentral hypergeometric distribution (see, for instance, Davison, 1988, Example 6.1). In order to obtain an integrated likelihood, we use here an idea suggested by Cox and Reid (1993), i.e., we choose a weight function based on the original parameterization that would act like a uniform one in an orthogonal parameterization, (ψ, ξ_i). Since the model is a full exponential family, (ψ, ξ_i) might be given by the so-called mixed parameterization. Hence, we have $|\partial\xi_i/\partial\lambda_i| = m_i e^{\lambda_i}/(1+e^{\lambda_i})^2 + e^{\psi+\lambda_i}/(1+e^{\psi+\lambda_i})^2$, which is a model-dependent weight function for the original parameter λ_i, which also depends on ψ. This leads to the integrated likelihood

$$L_O(\psi) = \prod_{i=1}^{q} \int \frac{e^{(y_{i1}+y_{i2})\lambda_i + y_{i2}\psi}}{(1+e_i^{\lambda})^{m_i+2}(1+e^{\psi+\lambda_i})^3} \{e_i^{\lambda}(1+e^{\psi+\lambda_i})^2 + e^{\psi+\lambda_i}(1+e_i^{\lambda})^2\} d\lambda_i.$$

After a change of variable $\lambda_i(\omega_i) = \log\{\omega_i/(1-\omega_i)\}$ and some algebra, we obtain

$$\begin{aligned} L_O(\psi) \quad = \quad & \prod_{i=1}^{q} e^{\psi y_{i2}} \left\{ {}_2F_1(1, y_{i1}+y_{i2}+1, m_i+2, 1-e^{\psi}) \right. \\ & \left. + e^{\psi}\ {}_2F_1(3, y_{i1}+y_{i2}+1, m_i+2, 1-e^{\psi}) \right\}, \end{aligned} \tag{4.1}$$

where ${}_2F(a,b,c,z)_1 = [\Gamma(c)/\{\Gamma(b)\Gamma(c-b)\}] \int_0^1 x^{b-1}(1-x)^{c-b-1}(1-zx)^{-a}dx$ (Abramowitz and Stegun, 1964, formula 15.3.1, page 558).

We also use the procedure based on the ZSE parameterization given by (2.1). Exploiting the exponential family framework, the new nuisance parameter ϕ_i is the solution of the implicit equation

$$K_{\lambda_i}(\hat{\psi}, \phi_i) - K_{\lambda_i}(\psi, \lambda_i) = 0, \tag{4.2}$$

where $K(\psi, \lambda_i) = m_i \log(1+e^{\lambda_i}) + \log(1+e^{\psi+\lambda_i})$ is the cumulant function, the subscript denotes the derivative with respect to λ_i, and $\hat{\psi}$ is the maximum

TABLE 1

Example 2. Empirical coverage (%) of R, R_C, R_I, R_O, and R_{MP} in three different settings: $m_i = 7$, $q = 300$ (top rows); $m_i = 5, 7, 9$, each replicated 100 times (middle rows); 30 strata with $m_i = 3$ and 270 strata with $m_i = 7$ (lower rows)

Nominal (%)	1.0	2.5	5.0	10.0	25.0	50.0	75.0	90.0	95.0	97.5	99.0
R	0.0	0.0	0.1	0.5	2.6	10.4	26.3	47.1	60.4	70.9	81.5
R_C	1.0	2.3	5.0	10.2	25.0	49.7	74.4	89.5	94.6	97.3	98.8
R_I	0.6	1.7	3.6	7.8	21.0	44.8	70.6	87.4	93.5	96.9	98.5
R_O	0.5	1.6	3.3	7.3	19.8	42.9	68.6	85.8	92.5	96.0	98.2
R_{MP}	0.6	1.8	3.7	8.0	21.1	44.8	70.2	87.1	93.2	96.5	98.3
R	0.0	0.1	0.2	0.6	2.8	9.9	25.4	46.1	59.7	70.4	80.7
R_C	1.0	2.8	5.1	10.2	25.1	49.8	75.1	90.2	94.7	97.3	99.0
R_I	0.8	2.1	4.0	8.0	20.5	44.4	70.4	87.5	93.4	96.5	98.5
R_O	0.6	1.9	3.8	7.6	19.8	43.2	69.4	86.5	92.9	96.0	98.2
R_{MP}	0.7	2.0	4.1	8.1	20.8	44.5	70.4	87.4	93.2	96.3	98.4
R	0.0	0.0	0.2	0.6	2.2	8.4	23.0	42.9	56.3	67.8	79.0
R_C	1.0	2.4	4.8	9.4	24.7	49.2	74.7	89.8	94.5	97.5	99.0
R_I	0.7	1.5	3.3	6.8	19.6	42.5	69.1	86.8	92.8	96.4	98.5
R_O	0.7	1.5	3.3	6.8	19.5	42.0	68.3	86.2	92.5	96.0	98.3
R_{MP}	0.7	1.6	3.5	7.2	20.3	43.3	69.6	86.8	92.8	96.2	98.4

likelihood estimate. Then we can obtain the integrated likelihood by a change of variable from ϕ_i to λ_i in the integrals,

$$\begin{aligned} L_I(\psi) &= \prod_{i=1}^{q} \int L_i(\psi, \lambda_i) \left| \frac{\partial \phi_i(\psi, \lambda_i; \hat{\psi})}{\partial \lambda_i} \right| d\lambda_i \\ &= \prod_{i=1}^{q} \int L_i(\psi, \lambda_i) \frac{K_{\lambda_i \lambda_i}(\psi, \lambda_i)}{K_{\lambda_i \lambda_i}(\hat{\psi}, \phi_i(\psi, \lambda_i; \hat{\psi}))} d\lambda_i. \end{aligned} \tag{4.3}$$

where the Jacobian $\partial \phi_i(\psi, \lambda_i; \hat{\psi}) / \partial \lambda_i$ is obtained by differentiating (4.2) with respect to λ_i. This Jacobian can be seen as a model and data-dependent weight function for the original parameter λ_i, and for this reason may depend on ψ. Note that the dependence on the data is only through $\hat{\psi}$. Of course we need ϕ_i as well in the integrand function; but for fixed λ_i, ψ and $\hat{\psi}$, it is possible to solve (4.2) numerically and get the corresponding ϕ_i. Finally, the integrals in (4.3) are computed numerically using adaptive quadrature (Piessens et al., 1983).

We perform some simulation studies, each based on 8,000 replications and with $\psi = \log(5)$, and λ_i equal to 1/8 plus a standard normal random noise. In particular, we consider three different settings with $q = 300$. The first setting has balanced strata with $m_i = 7$; the second setting has unbalanced strata, with m_i taking values 5, 7 and 9, each replicated 100 times. The average value of the m_i is the same as in the previous setting, i.e. $\bar{m} = 7$. Finally, the third setting is like the first one, but with 10% of the strata, i.e. 30, with a reduced sample size ($m_i = 3$), thus leading to $\bar{m} = 6.6$. Table 1 reports the empirical coverage probabilities of signed root likelihood ratio statistics based on profile likelihood (R), on conditional likelihood (R_C), on integrated likelihood with ZSE parameterization and uniform weight function (R_I), on (4.1) (R_O), and

TABLE 2
Example 2. Bias and root mean squared error (RMSE) of different estimators in three simulation settings: $m_i = 7$, $q = 300$ (left columns); $m_i = 5, 7, 9$, each replicated 100 times (middle columns); 30 strata with $m_i = 3$ and 270 strata with $m_i = 7$ (right columns)

Estimator	bias	RMSE	bias	RMSE	bias	RMSE
$\hat{\psi}$	0.250	0.320	0.258	0.329	0.275	0.343
$\hat{\psi}_C$	0.006	0.171	0.005	0.173	0.008	0.173
$\hat{\psi}_I$	0.026	0.170	0.028	0.173	0.035	0.174
$\hat{\psi}_O$	0.035	0.176	0.034	0.178	0.038	0.178
$\hat{\psi}_{MP}$	0.027	0.174	0.028	0.176	0.033	0.177

on modified profile likelihood (R_{MP}). Bias and root mean squared error of the corresponding estimators are reported in Table 2.

The empirical coverages of R_I and R_O are comparable, with the former being slightly more accurate and very close to the behavior of R_{MP}. Both R_I and R_{MP} give a reasonable approximation for R_C and improve substantially over R. The close agreement between R_I and R_{MP} can be explained by the results for exponential families in Severini (2007). The same indication can be found looking at bias and root mean squared error of the corresponding estimators in Table 2, although $\hat{\psi}_I$ has a smaller variance than $\hat{\psi}_{MP}$, thus leading to a reduced RMSE. The comments above apply equally to all three settings considered. On the other hand, the accuracy of the various methods is basically the same for the first two settings, i.e. those with the same average strata sample size $\bar{m} = 7$, while it is slightly worse in the last setting, probably due to the smaller amount of information ($\bar{m} = 6.6$).

4.3. First-order non stationary autoregressive model

Consider the first-order autoregressive model defined by

$$y_{ij} = \lambda_i + \rho y_{ij-1} + \varepsilon_{ij}, \tag{4.4}$$

where ε_{ij} are independent normal random variables with zero mean and variance σ^2, $j = 1, \dots, m_i$, $i = 1, \dots, q$.

When the time series in each stratum are stationary, that is when we assume $y_{i0} \sim N\{0, \sigma^2/(1-\rho^2)\}$, then λ_i is orthogonal to $\psi = (\rho, \sigma^2)$, and the modified profile likelihood and the integrated likelihoods proposed in this paper are equivalent and they all coincide with a marginal likelihood. The latter yields consistent estimates for ψ when q diverges, even for fixed m_i (Bartolucci et al., 2015, Example 1).

Here, we consider the non stationary case, which appears to be the dominant one in the econometric literature (see, for instance, Lancaster, 2002, Section 3). This means that we condition on the observed initial value y_{i0} and permit the autoregressive parameter to equal or exceed unity. Without loss of generality, in the following we will assume that $y_{i0} = 0$, $i = 1, \dots, q$. Indeed, this corresponds to assuming model (4.4) for the differences $y_{ij} - y_{i0}$, with λ_i reparameterized as $\lambda_i - y_{i0}(1-\rho)$. While the initial conditions have no influence in the definition of

the various methods presented below, they might have an effect on the amount of information in the sample; a detaild discussion is given in Dhaene and Jochmans (2014).

The log likelihood for model (4.4) is the sum of q independent components of the form

$$\ell^{(i)}(\rho, \sigma^2, \lambda_i) = -\frac{m_i}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{m_i}(y_{ij} - \lambda_i - \rho y_{ij-1})^2.$$

Lancaster (2002, page 655) shows that an information orthogonal parameterization is given by $\xi_i = \lambda_i \exp\{b(\rho)\}$, where

$$b_i(\rho) = \frac{1}{m_i}\sum_{j=1}^{m_i-1}\frac{m_i - j}{j}\rho^j. \tag{4.5}$$

The parameter ξ_i is orthogonal to both ρ and σ^2, with the latter two being orthogonal to each other.

Alternatively, we can use the ZSE parameterization (2.1), with ϕ_i solution of

$$E_{\rho_0,\sigma_0^2,\lambda_{i0}}\{\ell_{\lambda_i}(\rho, \sigma^2, \lambda_i)\}|_{(\rho_0,\sigma_0^2,\lambda_{i0})=(\hat\rho,\hat\sigma^2,\phi_i)} = 0. \tag{4.6}$$

Using again the results in Lancaster (2002), we find $\phi_i = \lambda_i/\{1 + (\hat\rho - \rho)b_i'(\hat\rho)\}$, where $b_i'(\rho) = (1/m_i)\sum_{j=1}^{m_i-1}(m_i - j)\rho^{j-1}$ is the first derivative of (4.5) and $\hat\rho$ is the maximum likelihood estimate. For this model computation of profile and integrated log likelihoods is straightforward since all maximization and integration involved can be easily done analytically. In particular, focusing interest on the parameter ρ, we have

$$\begin{aligned}
\ell_P(\rho) &= -\frac{\sum_{i=1}^q m_i}{2}\log SS(\rho), \\
\ell_O(\rho) &= -\frac{\sum_{i=1}^q (m_i - 1)}{2}\log SS(\rho) + \sum_{i=1}^q b_i(\rho), \qquad (4.7)\\
\ell_I(\rho) &= -\frac{\sum_{i=1}^q (m_i - 1)}{2}\log SS(\rho) - \sum_{i=1}^q \log\{1 + (\hat\rho - \rho)b_i'(\hat\rho)\}, \qquad (4.8)
\end{aligned}$$

where $SS(\rho) = \sum_{i=1}^q\sum_{j=1}^{m_i}\{w_{ij}(\rho) - \bar w_i(\rho)\}^2$, with $w_{ij}(\rho) = y_{ij} - \rho y_{ij-1}$, and $\bar w_i(\rho) = {m_i}^{-1}\sum_{j=1}^{m_i} w_{ij}(\rho)$. Formulae (4.7) and (4.8) are the integrated log-likelihoods with the orthogonal parameters ξ_i and with the ZSE parameters ϕ_i, respectively. In both cases we used a constant weight function for the incidental parameters and for $\log\sigma$. These integrated log likelihoods could also be obtained by first integrating out the incidental parameters, thus obtaining the integrated log likelihoods for (ρ, σ^2), and then profiling out σ^2.

Since the maximum likelihood estimate is generally highly biased, this could have an effect on the accuracy of the integrated likelihood (4.8). A possible solution could be given by using in (4.6) alternative estimates for ρ and σ^2 in place of $\hat\rho$ and $\hat\sigma^2$. One solution could be the use of a parametric bootstrap bias corrected version of $\hat\rho$ and $\hat\sigma^2$. Alternatively, one could use a different estimate,

such as for instance the maximizer of (4.7), or the maximizer of (4.8) itself, leading to a two-step solution. In the numerical example and in the simulations below we used the former option, thus obtaining the new ZSE parameter $\phi_i^I = \lambda_i/\{1 + (\hat{\rho}_O - \rho)b'(\hat{\rho}_O)\}$, where $\hat{\rho}_O$ denote the maximizer of (4.7). The corresponding integrated log-likelihood has the form (4.8), with $\hat{\rho}$ replaced by $\hat{\rho}_O$, and will be denoted by $\tilde{\ell}_I(\rho)$.

Sometimes $\ell_O(\rho)$ can be monotonic increasing for large values of ρ. On the other hand, for values of m_i, q and ρ of practical interest, it has a local maximum for $\rho \in (-\rho_l, \rho_u)$, where $\rho_l, \rho_u > 0$ are threshold values that can exceed one. Lancaster (2002), developing the integrated likelihood from a Bayesian perspective, shows that such a local maximum is a consistent estimator of ρ for large q, even for fixed m_i; see also Dhaene and Jochmans (2014). Also $\ell_I(\rho)$ and $\tilde{\ell}_I(\rho)$ can be monotonic increasing for large values of ρ, and this problem seems to occur "sooner" than for $\ell_O(\rho)$ for moderate values of m_i. On the other hand, for larger values of m_i this problem tends to disappear for $\ell_I(\rho)$ and $\tilde{\ell}_I(\rho)$, while it is accentuated for $\ell_O(\rho)$, given the polinomial form of the last term on the right hand side of (4.7). Instead, the second term in the right hand side of (4.8) cannot be computed for values of ρ greater than a certain threshold depending on $\hat{\rho}$ (which is however simple to determine and always greater than 1). A similar comment also applies to $\tilde{\ell}_I(\rho)$. Even in these cases, in practice, this has not proven to be a problem for maximization and inference.

As a numerical illustration, Figure 1 shows the relative log likelihoods for a simulated sample with $m_i = 8$, $i = 1, \ldots, 500$, $\rho = 0.9$, $\sigma^2 = 1$ and λ_i generated from a normal distribution with mean and variance equal to 1. The left panel shows the monotonicity issue for the integrated log likelihoods, while the right panel gives a zoomed version in an interval of values of practical interest for inference.

We also run some simulation studies, each with 10,000 simulated samples, $\rho = 0.9$, $\sigma^2 = 1$ and λ_i generated from a normal with mean and variance equal to 1, comparing the empirical coverage probabilities for the signed likelihood ratio statistics based on $\ell_P(\rho)$, $\ell_O(\rho)$, $\ell_I(\rho)$, and $\tilde{\ell}_I(\rho)$, which are denoted by R, R_O, R_I and $\tilde{R}_I$, respectively. Bias and mean squared errors of the corresponding estimators have also been considered. Tables 3 and 4 report the results for three different settings with $q = 500$. The first setting is the same as that used in the example of Figure 1, i.e. with balanced strata of size 8. The second setting has unbalanced strata, with m_i taking ten different values ranging from 3 to 12, each repeated 50 times. The average value of the m_i is the same as in the previous setting, i.e. $\bar{m} = 8$. The third setting is like the first one, but with 10% of the strata, i.e. 50, with a reduced sample size ($m_i = 4$), thus leading to $\bar{m} = 7.6$.

The results confirm the findings of Lancaster (2002) about $\ell_O(\rho)$. Indeed, this integrated likelihood provides consistent estimates and very accurate inference in all settings. On the other hand, $\ell_I(\rho)$, although largely improving over $\ell_P(\rho)$, does not have the same accuracy of $\ell_O(\rho)$. Such accuracy is better in the second setting with very unbalanced strata sample sizes. Instead, having a few strata with moderate sample size, as in the third setting, does not seem to lead to a significant loss of accuracy. Finally, we note that $\tilde{\ell}_I(\rho)$, while giving essentially

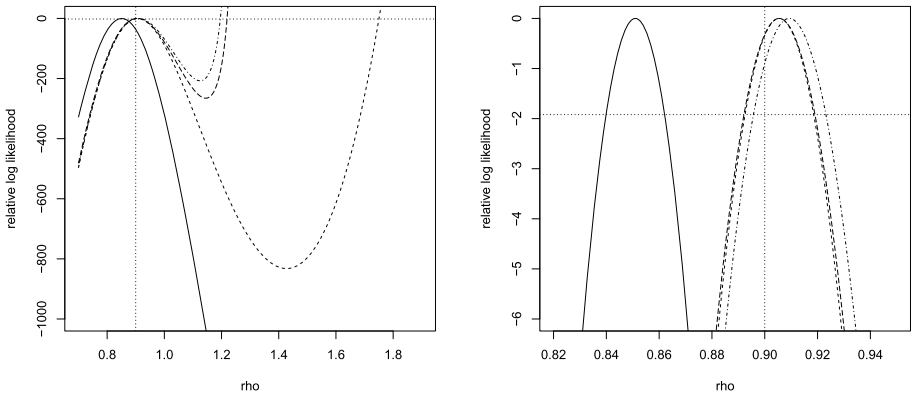

FIG 1. *Example 2. Relative log likelihoods for simulated data of the nonstationary autoregressive model:* $m_i = 8$, $q = 500$, $\rho = 0.9$, $\sigma^2 = 1$ *and* $\lambda_i \sim N(1, 1)$. *The solid line corresponds to* $\ell_P(\rho)$, *the dashed line to* $\ell_O(\rho)$, *the dot-dashed line to* $\ell_I(\rho)$, *and the long-dashed line to* $\tilde{\ell}_I(\rho)$. *The vertical dotted line indicates the true parameter value, while the horizontal dotted line provides confidence intervals of level 0.95 based on the corresponding likelihood ratio statistics. The left panel shows the uncostrained plot, while the right panel shows a zoomed version in a region of interest.*

TABLE 3

Example 3. Empirical coverage (%) of R, R_O, R_I *and* $\tilde{R}_I$ *in three simulation settings:* $m_i = 8$, $q = 500$ *(top rows);* $m_i = 3, 4, \ldots, 12$, *each replicated 50 times (middle rows);* 50 *strata with* $m_i = 4$ *and* 450 *strata with* $m_i = 8$ *(lower rows)*

Nominal (%)	1.0	2.5	5.0	10.0	25.0	50.0	75.0	90.0	95.0	97.5	99.0
R	100	100	100	100	100	100	100	100	100	100	100
R_O	1.1	2.7	5.5	10.7	25.7	49.8	74.2	89.2	94.3	97.0	98.6
R_I	0.2	0.6	1.4	3.4	11.5	29.4	54.7	75.8	85.4	91.3	95.7
$\tilde{R}_I$	0.9	2.4	4.9	10.0	25.0	49.8	74.9	90.0	95.0	97.4	98.9
R	100	100	100	100	100	100	100	100	100	100	100
R_O	1.0	2.7	5.1	10.1	25.1	49.9	74.0	89.0	93.9	96.9	98.6
R_I	0.4	1.0	2.2	5.0	15.2	35.8	62.1	81.2	89.3	93.7	97.1
$\tilde{R}_I$	0.8	2.3	4.7	9.6	24.6	50.0	74.5	89.5	94.4	97.2	98.9
R	100	100	100	100	100	100	100	100	100	100	100
R_O	1.2	3.1	5.8	11.2	26.3	49.9	74.3	89.1	94.2	96.9	98.7
R_I	0.2	0.7	1.6	3.6	11.6	29.2	53.7	75.4	85.0	90.9	95.4
$\tilde{R}_I$	1.0	2.7	5.3	10.4	25.5	49.8	75.0	89.9	94.8	97.3	99.0

TABLE 4

Example 3. Bias and root mean squared error (RMSE) of various estimators of ρ *in three simulation settings:* $m_i = 8$, $q = 500$ *(left columns);* $m_i = 3, 4, \ldots, 12$, *each replicated 50 times (middle columns);* 50 *strata with* $m_i = 4$ *and* 450 *strata with* $m_i = 8$ *(right columns)*

Estimator	bias	RMSE	bias	RMSE	bias	RMSE
$\hat{\rho}$	$-5.6 \cdot 10^{-2}$	0.0562	$-4.4 \cdot 10^{-2}$	0.0444	$-5.7 \cdot 10^{-2}$	0.0590
$\hat{\rho}_O$	$4.0 \cdot 10^{-5}$	0.0070	$1.2 \cdot 10^{-4}$	0.0064	$-1.4 \cdot 10^{-5}$	0.0074
$\hat{\rho}_I$	$4.0 \cdot 10^{-3}$	0.0084	$2.4 \cdot 10^{-3}$	0.0070	$4.4 \cdot 10^{-3}$	0.0090
$\tilde{\rho}_I$	$2.6 \cdot 10^{-5}$	0.0070	$1.1 \cdot 10^{-4}$	0.0064	$-2.7 \cdot 10^{-5}$	0.0074

the same estimates as $\ell_O(\rho)$, has $\tilde{R}_I$ which is slightly more accurate than R_O, in particular in the tails.

Results, not shown here, with $\rho = 0.5$ and/or with $q = 250$ show minor differences among the three integrated likelihoods. On the other hand, results in more extreme settings, such as with $\bar{m} = 4$ and $q = 1000$, qualitatively confirm the findings of the case with $q = 500$.

As a final remark, we note that the modified profile likelihood is not straightforward to obtain in this model. A possibility is to use the approximation of Severini (1998), avoiding the sometimes cumbersome analytical calculation of required expected value by means of Monte Carlo simulation. This approach is quite general, although computationally more intensive, and was also used by Bartolucci et al. (2015) for a dynamic regression model for binary data. Claudia Di Caterina, in an unpublished Master Thesis of the University of Padova, proved that Severini's approximation of the mixed derivative is linear in ρ. This implies that the modified profile log likelihood does not exist for certain values of ρ, similarly to $\ell_I(\rho)$. Moreover, it also shares the other drawbacks of $\ell_I(\rho)$, i.e., it could be monotonic increasing for not very large values of ρ, and the normal approximation for the corresponding signed likelihood ratio statistic has an accuracy very close to that of R_I, which is not very satisfactory for practical purposes when m_i is small and q is very large. We note that also the modified profile likelihood depends on the maximum likelihood estimates. Therefore it is possible that the use of better estimates could improve also its accuracy, as for the integrated likelihood $\tilde{\ell}_I(\rho)$, although, to our knowledge, this has not been investigated yet.

5. Discussion

In this paper we studied the frequentist asymptotic properties of the inferential quantities derived from integrated likelihoods in models with stratum nuisance parameters, in a two-index asymptotic setting with both sample size and number of nuisance parameters going to infinity. In particular, we showed that quantities based on an integrated likelihood, constructed using a properly chosen weight function, may have asymptotic behaviours close to the standard ones, and largely improving the accuracy of inference based on the corresponding quantities computed from the profile likelihood. Moreover, we showed that this kind of integrated likelihood has asymptotic properties similar to those of higher order methods such as the modified profile likelihood; the example in Section 4.1 shows that they can sometimes perform better in practice.

The weight function that guarantees good asymptotic properties is that applied to an orthogonal nuisance parameter, either based on the expected information or on the ZSE parameterization, and not depending on the parameter of interest. A constant weight function is typically a convenient choice, although both the theory and numerical examples, not shown here, indicate that the integrated likelihood is reasonably robust with respect to the chosen weight function. As shown in the example of Section 4.2, the construction of the integrated likelihood proposed here corresponds to using a model-dependent weight

function for the original nuisance parameter, in the case of expected information orthogonality, and a model and data-dependent weight function, in the case of ZSE orthogonality. Data-dependent weight functions for stratified models were also considered in Arellano and Bonhomme (2009). Although their proposal has some similarities with the one here, it seems less accurate, at least comparing simulation results for the binomial case.

The integrated likelihood is a tool for inference that can be applied in wide generality. On the other hand, the computation of integrals is required. This may seem a limitation, but the wide availability of numerical integration methods and the fact that the strata are independent allows one to parallelize the computation of many low dimensional integrals, thus increasing accuracy and substantially reducing the computational time.

Another possible issue is related to the need of an orthogonal parameterization in the construction of the integrated likelihood. Indeed, the information orthogonal parameterization may be not straightforward to compute or may not even exist for a vector parameter of interest. On the other hand, the ZSE parameterization can always be defined and has an algorithmic form that can be easily implemented, as shown in the example of Section 4.2.

The ZSE parameterization depends on the data through an estimate, typically the maximum likelihood estimate. The example of Section 4.3 shows that the use of alternative estimates may lead to more accurate inference. Another instance is given by an application of integrated likelihood in meta analysis, where the maximum likelihood estimate is often numerically unstable. Bellio and Guolo (2015) show that an integrated likelihood based on the ZSE parameterization constructed using an alternative estimate leads to very accurate inference. More work in this direction would be worthwhile.

We conclude with a comment on the assumptions of the two-index asymptotic setting. In Section 2 we assumed that the strata sample sizes m_i are asymptotically balanced, in the sense that they are of order $O(m)$ but not $o(m)$. This assumption implies that $\max_{1\le i\le q} m_i/n \to 0$ for $q \to \infty$, where $n = m_1 + \cdots + m_q$ is the total number of observations. However, the latter condition holds even if some strata have $m_i = o(m)$, provided the number of such strata does not increase with q. The results in the paper hold even under this weaker condition. The simulation results in the third asymptotic setting of both Sections 4.2 and 4.3 seem to confirm this claim. On the other hand, simulation results for the example in Section 4.2, not reported here, in the opposite scenario with many very small strata and only a few very large strata, indicate that it is not sufficient to have only a limited number of strata with a lot of information to guarantee a reasonable accuracy of the asymptotic approximations.

Acknowledgements

RDB was partially supported by grant BO3139/4-1 from the German Science Foundation (DFG). NS was partially supported by Progetto di Ateneo (CPDA131553), Università degli Studi di Padova. The work of TAS was supported by the National Science Foundation.

References

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications, New York.

Arellano, M. and Bonhomme, S. (2009). Robust priors in nonlinear panel data models. *Econometrica* **77** 489–536. MR2503036

Barndorff-Nielsen, O. E. (1996). Two order asymptotic. In *Frontiers in Pure and Applied Probability II: Proceedings of the Fourth Russian-Finnish Symposium Prob. Th. Math. Statist.* (A. Melnikov, ed.) 9–20. TVP Science, Moscow.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*. Chapman and Hall, London. MR1317097

Bartolucci, F., Bellio, R., Salvan, A. and Sartori, N. (2015). Modified profile likelihood for fixed effects panel data models. *Econometric Reviews*, to appear.

Bellio, R. and Guolo, A. (2015). Integrated likelihood inference in small sample meta-analysis. *Scandinavian Journal of Statistics*, to appear.

Berger, J. O., Liseo, B. and Wolpert, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science* **14** 1–22. MR1702200

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)* **49** 1–39. MR0893334

Cox, D. R. and Reid, N. (1993). A note on the calculation of adjusted profile likelihood. *Journal of the Royal Statistical Society. Series B (Methodological)* **55** 467–471. MR1224410

Davison, A. C. (1988). Approximate conditional inference in generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)* **50** 445–461. MR0970979

Dhaene, G. and Jochmans, K. (2014). Likelihood inference in an autoregression with fixed effects. *Econometric Theory* **FirstView** 1–38.

Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika* **80** 3–26. MR1225211

Evans, M. and Swartz, T. (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford University Press, New York. MR1859163

Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society. Series B (Methodological)* **32** 175–208. MR0270474

Lancaster, T. (2002). Orthogonal parameters and panel data. *Review of Economic Studies* **69** 647–666. MR1925308

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference: From a Neo-Fisherian Perspective*. World Scientific, Singapore. MR1476674

Piessens, R., de Doncker-Kapenga, E., Überhuber, C. and Kahaner, D. (1983). *Quadpack: A Subroutine Package for Automatic Integration*. Springer Verlag, Berlin. MR0712135

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (2007). *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, New York. MR0833288

Reid, N. (2003). Asymptotics and the theory of inference. *Annals of Statistics* **21** 1695–1731. MR2036388

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika* **90** 533–549. MR2006833

Severini, T. A. (1998). An approximation to the modified profile likelihood function. *Biometrika* **85** 403–411. MR1649121

Severini, T. A. (2000). *Likelihood Methods in Statistics*. Oxford University Press, New York. MR1854870

Severini, T. A. (2007). Integrated likelihood functions for non-Bayesian inference. *Biometrika* **94** 529–542. MR2410006

Severini, T. A. (2010). Likelihood ratio statistics based on an integrated likelihood. *Biometrika* **97** 481–496. MR2650752

Severini, T. A. (2011). Frequency properties of inferences based on an integrated likelihood function. *Statistica Sinica* **21** 433–447. MR2796870

Sweeting, T. J. (1987). Discussion of the paper by Cox and Reid. *Journal of the Royal Statistical Society. Series B (Methodological)* **49** 20–21.

Ventura, L., Cabras, S. and Racugno, W. (2009). Prior distributions from pseudo-likelihoods in the presence of nuisance parameters. *Journal of the American Statistical Association* **104** 768–774. MR2541593

Wooldridge, J. M. (1994). Estimation and inference for dependent processes. *Handbook of Econometrics* **4** 2639–2738. MR1315980