SMGroup

SM Journal of Biometrics & Biostatistics

Review Article

On the Choice of the Weight Function for the Integrated Likelihood

Alessandro Casa*

Department of Statistical Sciences, University of Padova, Italy

Article Information

Received date: Jul 27, 2018
Accepted date: Aug 06, 2018
Published date: Aug 16, 2018

***Corresponding author**

Alessandro Casa, Department of Statistical Sciences, University of Padova, Italy, Tel: (0039) 049 8274111; Email: casa@stat.unipd.it

Keywords Bayesian Inference; Likelihood Inference; Nuisance Parameter; Profile Likelihood; Pseudo Likelihood; Reference Prior

Abstract

In the field of biostatistics it is often required to develop inferential tools dealing with the presence of nuisance parameters. The most adopted solution is to resort to pseudo-likelihood functions, having properties similar to the ones of a genuine likelihood. A possible choice is to use the integrated likelihood where the nuisance parameters are eliminated by means of integration with respect to a weight function. The selection of the weight function turns out to be crucial since it could have a strong impact on the properties of the resulting integrated likelihood. After having introduced the concept of pseudo-likelihood, the definition and the properties of the integrated likelihood, the focus will be on reviewing the main alternatives to choose the weight function according to different inference paradigms.

Introduction

Statistical inference procedures aim to gain knowledge about a phenomenon of interest. In a parametric context, where $F=\{p_Y(y;\theta), y\in Y, \theta\in\Theta\}$ is the considered model, this corresponds to have indications about plausible values for the parameter θ. In real applications, as for instance those in biomedical field, θ is often partitioned as $\theta=(\psi,\lambda)$ where $\psi\in\Psi$ and $\lambda\in\Lambda$ are respectively the parameter of interest and the nuisance parameter, both possibly multidimensional. Nuisance parameters are not of primarily concern, nervertheless sometimes they are needed to have a more realistic modelling and representation of the phenomenon under study. On the other hand, their presence introduces some relevant issues when considering inferential procedures. For this reason, the problem of eliminating nuisance parameters has become a central one in statistical literature and has been faced adopting several different perspectives.

In a frequentist framework the main tool used to carry out inferential procedures is the likelihood function $L(\theta)$; for a comprehensive treatment of its properties and characteristics see Pace & Salvan [1]. When dealing with the presence of nuisance parameters, to gain better inferential performances, we need to isolate interesting features of the likelihood function.

A convenient way to proceed is to resort to conditional or marginal likelihood (see, e.g., Pace & Salvan, 1997, Ch.4), [1]. They are both genuine likelihoods, respectively under the conditional and marginal reduced models, therefore sharing the same properties of the likelihood function. Unluckily, marginal and conditional likelihood arise essentially only in some specific situations, e.g. when dealing with exponential and group families, thus reducing their applicability in more complex situations.

Alternatively several ways to obtain pseudo-likelihood functions have been proposed. Pseudo-likelihood is a function whose behavior is as similar as possible to the one of a genuine likelihood but which is broadly based on an incomplete specification of the underlying model. In the current context it has to be intended as a function of data depending only on the parameter of interest. More formally, let $\Lambda(\psi)=\{\lambda:(\psi,\lambda)\in\Theta\}$ denote the parameter space for λ when ψ is fixed and $L_\psi=\{L(\psi,\lambda):\lambda\in\Lambda(\psi)\}$ the corresponding set of likelihood functions. The idea, in the derivation of pseudo-likelihood, is to build a function summarizing L_ψ. However, as noted by Severini [2], the construction of pseudo-likelihood involves some arbitrariness in defining what constitutes an effective summary of L_ψ; this has led to different proposals, having different rationales behind.

The profile likelihood is one of the most used pseudo-likelihoods. It summarizes L_ψ using its maximum value as $L_p(\psi)=L(\psi,\hat{\lambda}_\psi)=\sup_{\lambda\in\Lambda(\psi)}L(\psi,\lambda)$. Despite being widely used, profile likelihood has some well known drawbacks encountered for example when dealing with particular shapes of the likelihood function [3], or when the number of nuisance parameters is large with respect to the sample size. To overcome some of these drawbacks several modifications have been proposed as the modified profile likelihood [4] and the adjusted profile likelihood [5]. These modifications are known to perform well in the elimination of nuisance parameters and to have some desiderable properties as, for example, they approximate marginal or conditional likelihoods, when available.

This work focuses on a different approach to eliminate nuisance parameters by the construction of a pseudo-likelihood, that is the integrated likelihood [2,3,6,7]. The integrated likelihood summarizes L_ψ by mean of integration with respect to the so called weight function $\pi(\lambda|\psi)$. As will be clarified in the next section there are some advantages in using the integrated likelihood when dealing with nuisance parameters but, at the same time, the selection of the weight function $\pi(\lambda|\psi)$ is required. Generally there is not a straightforward way to select the weight function, with the exception for the case of group families where the right invariant Haar density for λ should be considered [8]. At the same time this choice could have a strong impact on the resulting integrated likelihood and on the derived inferential procedures. The aim of this work is to review the proposed methods to select the weight function.

In Section 2 we briefly introduce the integrated likelihood and its properties. In Section 3 we examine some proposed methods to select $\pi(\lambda|\psi)$ trying to take into consideration both the frequentist and the Bayesian paradigms. Lastly, in Section 4, we present some concluding remarks.

Integrated Likelihood

The integrated likelihood is defined as

$$L_I(\psi) = \int_\Lambda L(\psi,\lambda)\pi(\lambda|\psi)d\lambda$$

and constitutes an interesting way to deal with the presence of nuisance parameters. The idea is to build a summary of L_ψ substituting the maximization step, required in the derivation of the profile likelihood, with an average. This leads to several advantages: first of all, being an average over all the conditional likelihood functions given nuisance parameters, it allows incorporating directly the corresponding uncertainty. A maximization approach does not allow to take into account this uncertainty and many of the available modifications of the profile likelihood are indeed adjustments aiming to consider this drawback. This implies that $L_I(\psi)$ offers often a more informative summary of L_ψ with respect to $L_p(\psi)$ also, the asymptotic properties of the integrated likelihood are superior with respect to the ones of the profile likelihood. Moreover, the integrated likelihood is always available, also in non-regular and more complex models where conditional and marginal likelihood cannot be used. Lastly, besides the theoretical motivations, computationally the maximization step could be trickier with respect to the integration that is often numerically more stable and for which we can rely on several methods specifically conceived for solving integrals: Grazian & Liseo [9] show how to obtain an integrated likelihood considering ABC algorithms while, for a detailed review about integration algorithms to compute the integrated likelihood, the reader could refer to Zhao & Severini [10].

The integrated likelihood naturally comes out in a Bayesian framework where the weight function is the conditional prior density of λ given ψ. As Berger et al. [3] point out; the elimination of nuisance parameters in a Bayesian setting does not attract too much attention because it seems obvious to consider, for example, a uniform-integrated likelihood as

$$L_u(\psi) = \int_\Lambda L(\psi,\lambda)d\lambda$$

where a uniform prior for λ is considered.

In this framework the choice of the weight function could naturally be recast as a prior elicitation problem, a topic that has attracted a great amount of literature that cannot be comprehensively reviewed here. The immediacy of this approach could be somehow lost when facing situations with a great number of nuisance parameters: since it is possible that these parameters do not have a clear meaning, the elicitation of the corresponding prior distributions could be troublesome.

The use of integrated likelihood has also been studied when considering a frequentist approach to inference. It has been shown that, with specific choice for the weight function, the resulting pseudo-likelihood could have some properties of a genuine likelihood. A relevant difference with Bayesian framework is that, in this context, we do not have to worry about $\pi(\lambda|\psi)$ being a proper density function.

The next section aims to review some of the proposed methods to select the weight function, highlighting properties, advantages and drawbacks of each one of them.

Choice of the Weight Function

Bayesian approaches

In Liseo [7] we can find one of the first attempts to analyze the effect of different choices of the weight function on the resulting integrated likelihood. The elimination of nuisance parameters is studied from a Bayesian point of view, highlighting how the integrated likelihood comes out naturally in this framework. Let $\pi(\psi,\lambda) = \pi(\psi)\pi(\lambda|\psi)$ be the joint prior distribution. Inference on the parameter of interest is based on its posterior distribution (i.e. the marginal posterior distribution) that, given a sample $y = (y_1,\ldots,y_n)$ is defined as

$$\pi(\psi|y) \propto \pi(\psi)\int_\Lambda L(\psi,\lambda)\pi(\lambda|\psi)d\lambda = \pi(\psi)L_I(\psi),$$

clearly showing how integrated likelihood obviously emerges.

Therefore, from a Bayesian point of view, the critical step is given by the elicitation of the prior distribution. Since often we do not have previous information about nuisance parameters, frequently an objective approach has been considered, trying not to introduce subjective information about parameters that are not of interest.

Among the available choices of noninformative prior distributions the Jeffreys one [11] and the *reference prior* [12] are commonly used.

In the unidimensional case the *Jeffreys prior* for a parameter is defined ν as

$$\pi_J(\nu) \propto \sqrt{i(\nu)},$$

where $i(\nu) = -E\left\{\partial^2\ell(\nu)/\partial\nu^2\right\}$ is the Fisher information with $\ell(\nu) = \log L(\nu)$. This prior is invariant under reparametrization and does not depend on the declared parameter of interest. In multidimensional cases Jeffreys itself questioned the effectiveness of this approach since this method would seek the noninformative prior for the entire vector of parameters introducing a form of bias in the procedure, when the interest is restricted only to some elements of the vector. As an alterntive, Jeffreys proposed to assume

independence between the parameters and to obtain the joint prior distribution as the product of the marginal ones. This way to proceed creates a link with the orthogonalization idea that will be introduced when considering frequentist approaches in the following.

The reference prior method could be seen as a way to overcome the issues of the Jeffreys' one in multidimensional context. It is based on the maximization of an asymptotic version of the expected Kullback-Leibler divergence information between the prior and the posterior parameter distribution. The idea is to use the least informative prior distribution in the sense of letting the data to "speak most loudly"; for a more formal definition see Berger et al. [13]. In the specific framework considered, an appealing characteristic of the reference prior method is that, in multiparameter case, it allows for an explicit distinction between parameters of interest and nuisance ones inducing an importance ordering and grouping of the parameters according to the specific inferential aim.

When dealing with the elimination of nuisance parameters Liseo [7] shows that, in some particular examples of interest, the use of the reference prior leads to better performances with respect to Jeffreys and classical frequentist approaches such as profile likelihood and its modifications.

As Berger et al. [3] state "there are as many integrated likelihood as there are priors"; thus since subjectivity could have its role in the prior elicitation, it turns out to be impossible to list comprehensively the possible choice for the weight function in this framework. A promising and interesting choice that is worth citing consists in consider probability matching priors [14-17]. Here the idea is to use a prior distribution ensuring either exact or approximate validity of Bayesian credible regions from a frequentist point of view. Further alternative ways to obtain a marginal posterior distribution in a Bayesian setting can be found e.g. in Kass et al. [18] and Leonard et al. [19].

Lastly note that, even if their attention is not restricted to this framework, a more thorough discussion about the use of integrated likelihood in Bayesian framework can be found in Berger et al. [3].

Frequentist approaches

Although it arises in an intuitive manner in a Bayesian framework, the integrated likelihood can be used effectively also in a frequentist context and some attempts to link the choice of the weight function to the usefulness of the integrated likelihood for non-Bayesian inference have been made. Methods to select $\pi(\lambda|\psi)$ have been developed following different routes.

One approach attempts to build an integrated likelihood having some of the properties of a genuine likelihood. The seminal work by Severini [2] studies how to select $\pi(\lambda|\psi)$ in such a way that the resulting integrated likelihood behaves as a genuine likelihood. A first remark is that, since there is arbitrariness in the choice of the weight function, $L_I(\psi)$ should be as insensitive as possible to it; we find this to be coherent with the consideration in Liseo [7] about the adequacy of noninformative priors.

Before going through the details of the proposed method it is necessary to introduce the concept of *weakly unrelated* and *strongly unrelated* parameter. A parameter λ is weakly unrelated to ψ if $\hat{\lambda}_\psi = \hat{\lambda} + 0(n^{-1})$ for deviation of ψ such that $\psi = \hat{\psi} + 0\left(n^{-\frac{1}{2}}\right)$ where $\hat{\lambda}_\psi$ is the maximum likelihood estimator of λ for fixed ψ and $\hat{\psi}$ is the maximum likelihood estimator of ψ. On the other hand, two parameters are strongly unrelated if they are weakly unrelated and, in addition, $\hat{\lambda}_\psi = \hat{\lambda} + 0(n^{-\frac{1}{2}})$ for deviation of ψ such that $\psi - \hat{\psi} = 0(1)$. Note that orthogonality of ψ and λ assures that parameters are weakly unrelated, while it is not sufficient for the stronger condition (see e.g. [1], Ch. 4.7).

If a nuisance parameter λ is, or can be derived as at least weakly unrelated to the parameter of interest ψ, Severini [2] suggests to choose $\pi(\lambda|\psi)$ in such a way that ψ and λ are independent. The resulting likelihood will have indeed the following desiderable properties:

- Suppose that the likelihood can be factorize as $L(\theta) = L_1(\psi)L_2(\lambda)$. Choosing the weight function as suggested assure that the resulting integrated likelihood will correspond to $L_1(\psi)$, giving an intuitive result;
- When considering an integrated likelihood $L_I(\psi)$, generally the first two Bartlett identities do not hold. Severini [2] shows that, if we choose the weight function in the proposed way, $E\{\ell'_I(\psi);\theta\} = 0(n^{-1})$ where $\ell_I(.) = \log L_I(.)$ and the superscripts indicates the number of derivatives to be taken. Furthermore, if λ and ψ are strongly unrelated, $E\{\ell''_I(\psi) + \ell'_I(\psi)\ell'_I(\psi)^T;\theta\} = 0(n^{-1})$. Therefore this proposal allows to recover the score and information unbiasedness, at least asymptotically;
- if we consider a Laplace approximation of the integrated likelihood we get

$$L_I(\psi) = cL_p(\psi)\left|-\ell_{\lambda\lambda}(\psi,\hat{\lambda}_\psi)\right|^{-\frac{1}{2}}\pi(\hat{\lambda}_\psi|\psi)\{1+0(n^{-1})\}, \qquad (1)$$

  with $\ell_{\lambda\lambda}(\psi,\lambda) = \partial^2\ell(\psi,\lambda)/\partial\lambda^2$ and c not depending on ψ. Again, if we work with strongly unrelated parameters and if we select the weight function in the way proposed, we will have that, ignoring terms of order $0(n^{-\frac{1}{2}})$, $L_I(\psi)$ will not depend on the form of $\pi(\lambda)$. This turns out to be relevant since we want to work with integrated likelihood being insensitive to the features of the prior density for the nuisance parameter;
- The resulting integrated likelihood will be invariant with respect to interest-respecting reparametrization of the form $\tau = \tau(\psi,\lambda)$ with $\tau = (\tau_1,\tau_2)$ such that $\tau_1 = \tau_1(\psi)$ and $\tau_2 = \tau_2(\psi,\lambda)$.

Lastly, resorting again to the Laplace approximation given in (1), it can be shown (see [2], Appendix 2) that $L_I(\psi) = c_0\bar{L}_M(\psi)\{1+0(n^{-\frac{1}{2}})0(|\psi-\hat{\psi}|)\}$, where c_0 is a constant not depending on ψ and $\bar{L}_M(\psi)$ is an approximation of the modified profile likelihood of Barndorff-Nielsen [4]. This relation is useful both to study the properties of $L_I(\psi)$ and to further highlight why the way proposed by Severini [2] to choose the weight function could be particularly appropriate since the modified profile likelihood is known to have some desiderable properties and to work well in many applications.

A possible limitation of this approach is the need to have unrelated parameters. It is known that we can obtain weakly unrelated parameters starting from the orthogonal parametrization but this

requires the solution of differential equations that can be tricky. Furthermore, if ψ is not a scalar, the solution may not exist (see, e.g. Cox & Reid, 1987). In order to avoid such complications, Severini [2] proposes a way to construct a nuisance parameter strongly unrelated to the one of interest. In specific, let $\phi \equiv \phi(\psi,\lambda;\psi)$ be the zero-score-expectation parameter obtained as the solution to the equation

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

It can be shown that ϕ is strongly unrelated to ψ. When studying frequentist properties of the integrated likelihood we have to take into account that ϕ, defined as before, depends on the observed data.

A further, alternative approach aims at constructing the integrated likelihood in order to result in superior statistical procedures. It finds its motivation considering that the statistical procedures based on the integrated likelihood derived as in Severini (2) are not necessarily superior to those obtained choosing the weight function in other ways. The aim, when obtaining pseudo-likelihood, is to have a useful tool for subsequent inferential goals. Therefore is certainly interesting to study if it is possible to obtain, choosing appropriately $\pi(\lambda|\psi)$ an integrated likelihood giving better inferential performances.

Severini [20] starting from these considerations focuses on studying the properties of the resulting inferential procedures rather than on studying the properties of the integrated likelihood itself. Assume ψ to be scalar. We define the signed likelihood ratio statistic as

$$R = \operatorname{sgn}\left(\hat{\psi}-\psi\right)\left[2\left\{\ell_p\left(\hat{\psi}\right)-\ell_p(\psi)\right\}\right]^{\frac{1}{2}},$$

while the integrated likelihood ratio statistic is simply obtained by substituting ℓ_p, the profile log-likelihood, with the corresponding integrated log-likelihood, giving

$$\overline{R} = \operatorname{sgn}\left(\overline{\psi}-\psi\right)\left[2\left\{\ell_I\left(\overline{\psi}\right)-\ell_I(\psi)\right\}\right]^{\frac{1}{2}},$$

where $\overline{\psi}$ is the maximizer of $\ell_I(\psi)$.

The main aim of the work is to study the asymptotic properties of $\overline{R}$. As a side and related contribution the impact of the weight function on these properties is studied. For example it is shown that, choosing the weight function in the way introduced in the previous section, we have that $\overline{R}$ is standard normal with error $0\left(n^{-1}\right)$. Another possibility is given by choosing $\pi(\lambda|\psi)$ in such a way that

$$\frac{\partial}{\partial\psi}\log\pi(\lambda|\psi) = \frac{1}{6}\frac{\mu_{\psi,\psi,\psi}(\psi,\lambda)}{\bar{i}_{\psi,\psi}(\psi,\lambda)} + 0\left(n^{\frac{1}{2}}\right),$$

where $\mu_{\psi,\psi,\psi} = E\left\{\ell_\psi(\psi,\lambda)^3;\psi,\lambda\right\}/n$ and $i_{\psi,\psi}(\psi,\lambda) = E\left\{\ell_\psi(\psi,\lambda)^2;\psi,\lambda\right\}/n$. Thus it is assured that the resulting integrated likelihood is asymptotically standard normal to second order.

The work by Severini [20] allows to understand how the choice of $\pi(\lambda|\psi)$, affecting the properties of $L_I(\psi)$, could obviously have also an impact on the inferential process and how we can therefore choose it in such a way to gain better asymptotic properties for inferential procedures.

In Severini [21] the focus is still on choosing the weight function aiming to have resulting statistical procedures based on the integrated likelihood that enjoys some good properties. It is specifically highlighted that score and information unbiasedness, not being directly involved in the construction of statistical procedures, do not automatically guarantee optimal statistical properties. Hence the author focuses on the coverage probabilities of the integrated likelihood ratio based confidence intervals and on the mean squared error of the maximum integrated likelihood estimator.

Considering a Laplace approximation $L_I(\psi)$ can be expressed as

$$L_I(\psi) = L_A(\psi)\pi\left(\hat{\lambda}_\psi\middle|\psi\right)\left[1+0\left(n^{-1}\right)\right], \qquad (2)$$

where $L_A(\psi)$ is the Cox-Reid adjusted profile likelihood [5]. From this relation it can be shown that, if the aim is to work with an approximately score unbiased integrated likelihood, $L_A(\psi)$ could be used. Therefore the subsequent focus is on showing how an adequate choice of the weight function could lead to an integrated likelihood having better frequency properties with respect to $L_A(\psi)$.

Define $h(\psi,\lambda) = \log\pi(\lambda|\psi), h_\psi(\psi,\lambda) = \partial h(\psi,\lambda)/\partial\psi$ and let $i_{\psi,\psi}$ be the block of the expected information matrix for the parameter of interest. Furthermore let C_I and C_A be respectively the length of the likelihood ratio confidence interval based on $L_I(\psi)$ and $L_A(\psi)$.It can be shown that the expectation of C_I will be smaller with respect to the one of C_A whenever $h_\psi(\psi,\lambda)/i_{\psi\psi}(\psi,\lambda)$ is decreasing in ψ. Therefore we will have shorter intervals if $\partial\left[h_\psi(\psi,\lambda)/i_{\psi\psi}(\psi,\lambda)\right]/\partial\psi$ is sufficiently negative.

Analogous results are found when considering the mean squared error of the point estimators based on $L_I(\psi)$ and $L_A(\psi)$. The author shows indeed that $\hat{\psi}_I$ is again preferable to $\hat{\psi}_A$ whenever $\partial\left[h_\psi(\psi,\lambda)/i_{\psi\psi}(\psi,\lambda)\right]/\partial\psi$ is sufficiently negative. Note that these results also hold considering the approximation to the modified profile likelihood considered in previous sections. Lastly a word of caution is needed since the discussed results are not invariant with respect to reparametrizations of the parameter of interest hence we should carry on the comparison keeping in mind the parametrization used.

The reason why we obtain the results above is that the prior density in the integrated likelihood imposes a stochastic constraint on the parameters. We can see the approximation in (2) as if we were adding to $L_A(\psi)$ an observation $\hat{\lambda}$ having weight given by $\pi(\hat{\lambda}|\psi)$. Since the density of $\hat{\lambda}$ should not be $\pi(.|\psi)$, the addition leads to an improvement whenever $\pi(\hat{\lambda}|\psi)$ is in accordance with the true density. Thus choosing adequately the weight function, we can have improved performances with respect to the ones obtainable focusing on Bartlett's identities.

Miscellanea

In the previous sections we focused on two different approaches to the problem of choosing the weight function in the integrated likelihood by discussing some of the most relevant works in these frameworks since providing a complete review of the literature is beyond the scope of the paper. Nonetheless as we said the problem of choosing a prior, from a Bayesian point of view, has led to a great amount of literature both from an objective and a subjective perspective. The selection of the weight function could also be based on some subjective reasons based on possible knowledge on the nuisance parameter in specific applications. For example Kitakado et

al. [22] base their choice of the prior density on some beliefs related to the specific phenomenon under study and hence, in this specific case, linked to some previously available genetic knowledge.

The issue of choosing a weight function has been faced also in model with stratum nuisance parameters [23]. Let be $y=(y_1,...,y_q)$ a sample where y_i is a realization from the m_i-dimensional random variable Y_i having density $p_i(y_i;\psi,\lambda_i)$ where ψ is the parameter of interest and $\lambda=(\lambda_1,...,\lambda_q)$ the nuisance one. It is not unlikely that the number of nuisance parameters is of the same order as the sample size and, in this situation; procedures based on profile likelihood are known to perform poorly. Assuming independence among the strata the likelihood can be expressed as $L(\psi,\lambda)=\prod_{i=1}^{q}L^{(i)}(\psi,\lambda_i)$, where $L^{(i)}(\psi,\lambda_i)$ constitutes the likelihood contribution for the i-th stratum and q is the sample size as well as the dimension of λ. Hence the integrated likelihood in this setting is defined as $L_I(\psi)=\prod_{i=1}^{q}\int_{\Lambda}L^{(i)}(\psi,\lambda_i)\pi(\lambda_i|\psi)d\lambda_i$. It is shown that choosing the weight function as in Severini [2], with adjustments due to the particular context, leads to an improvement in the accuracy of inference with respect to profile likelihood based solutions. Moreover the resulting properties of the integrated likelihood are similar to those of modified profile likelihood.

One of the possible drawbacks of the previous approach is that, to obtain the integrated likelihood, the model has to be reparametrized. In literature, in the same framework, there are some proposed alternatives such as the one in Arellano and Bonhomme [24]. The authors propose data-dependent priors allowing not to reparametrize the model and show that this priors still are able to reduce the bias of the score function. Lastly note that this approach shares a strong link to the one proposed in Severini [2] if a reparametrization of unrelated parameters is considered.

Other works resorting to the integrated likelihood to eliminate nuisance parameters can be found in various areas. For example Cortese & Sartori [25] consider a similar problem as in De Bin et al. [23] but dealing with survival models with clustered and censored data. In He & Severini [26] the authors assume a Gaussian process as weight function to average over the unknown function in a semiparametric regression model setting. Bellio & Guolo [27] show that an integrated likelihood approach could be good also for small sample inference in Meta analysis leading to better inferential properties with respect to the modified profile likelihood. Lastly this approach has found fruitful application also when modelling data coming from capture-recapture type experiments [28].

Discussion and Conclusion

When dealing with inferential problems in the presence of nuisance parameters a common solution, to avoid the related drawbacks, is to resort to pseudo-likelihood functions. An appealing choice, in this context, is represented by the integrated likelihood. Some of the reasons why it would be preferable to consider the integrated likelihood over the profile likelihood have been highlighted in the article. Despite these advantages, this approach requires the use of a weight function that could have an impact on the properties of the resulting pseudo-likelihood and whose choice is not straightforward.

In this work we reviewed some of the possible way to proceed when choosing the weight function. Integrated likelihood arises naturally in Bayesian paradigm to inference but could be successfully considered also from a frequentist point of view. We tried to consider both these approaches highlighting the main differences and similarities.

Generally speaking the works in the literature do not give indications about a specific weight function but allow selecting a class of functions having some characteristics. It is relevant to point out that there is not an optimal choice in selecting $\pi(\lambda|\psi)$ since the definition of optimality could depend on the specific aim of the analysis. Concluding, note also that in the analysis certain formulations of the considered models could suggest a convenient choice for the weight function; see Berger et al. [3] for a more detailed discussion.

References

1. Pace L, Salvan A. Principles of statistical inference from a neo-fisherian perspective. Singapore: World Scientific Publishing Co. 1997.
2. Severini TA. Integrated likelihood functions for non-bayesian inference. Biometrika. 2007; 94: 529-542.
3. Berger JO, Liseo B, Wolpert RL. Integrated likelihood methods for eliminating nuisance parameters. Statistical Science. 1999; 14: 1-28.
4. Barndorff-Nielsen O. On a formula for the distribution of the maximum likelihood estimator. Biometrika. 1983; 70: 343-365.
5. Cox DR, Reid N. Parameter orthogonality and approximate conditional inference. Journal of the Royal Statistical Society. Series B (Methodological). 1987; 49: 1-39.
6. Kalbfleisch JD, Sprott DA. Application of likelihood methods to models involving large numbers of parameters. Journal of the Royal Statistical Society. Series B (Methodological). 1970; 32: 175-208.
7. Liseo B. Elimination of nuisance parameters with reference priors. Biometrika. 1993; 80: 295-304.
8. Berger JO. Statistical decision theory and bayesian analysis. New York: Springer Series in Statistics. 1985.
9. Grazian C, Liseo B. Approximate integrated likelihood via abc methods. Statistics and Its Interface. 2016; 2: 161-171.
10. Zhao Z, Severini TA. Integrated likelihood computation methods. Computational Statistics. 2017; 32: 281-313.
11. Jeffreys H. Theory of probability. Oxford University Press. 1961.
12. Berger JO, Bernardo J. On the development of the reference prior method. Bayesian Statistics. 1992; 4: 35-60.
13. Berger JO, Bernardo JM, Sun D. The formal definition of reference priors. The Annals of Statistics. 2009; 37: 905-938.
14. Gosh JK, Rahul M. Noninformative priors. Bayesian Statistics. 1992; 4: 195-203.
15. Berger JO, Philippe A, Robert CP. Estimation of quadratic functions: noninformative priors for non-centrality parameters. Statistica Sinica. 1998; 8: 359-375.
16. Datta GS, Sweeting TJ. Probability matching priors. Handbook of statistics. 2005; 25: 91-114.
17. Staicu AM, Reid NM. On probability matching priors. Canadian Journal of Statistics. 2008; 36: 613-622.
18. Kass RE, Tierney L, Kadane JB. A symptotics in bayesian computation. Bayesian Statistics. 1988; 3:261-278.
19. Leonard T, Hsu JS, Tsui KW. Bayesian marginal inference. Journal of the American Statistical Association. 1989; 84: 1051-1058.
20. Severini TA. Likelihood ratio statistics based on an integrated likelihood. Biometrika. 2010; 97: 481-496.

21. Severini TA. Frequency properties of inferences based on an integrated likelihood function. Statistica Sinica. 2011; 21: 433-447.

22. Kitakado T, Kitada S, Kishino H, Skaug HJ. An integrated-likelihood method for estimating genetic differentiation between populations. Genetics. 2006; 173: 2073-2082.

23. De Bin R, Sartori N, Severini TA. Integrated likelihoods in models with stratum nuisance parameters. Electronic Journal of Statistics. 2015; 9: 1474-1491.

24. Arellano M, Bonhomme S. Robust priors in nonlinear panel data models. Econometrica. 2009; 77: 489-536.

25. Cortese G, Sartori N. Integrated likelihoods in parametric survival models for highly clustered censored data. Lifetime data analysis. 2016; 22: 382-404.

26. He H, Severini TA. A flexible approach to inference in semiparametric regression models with correlated errors using gaussian processes. Computational Statistics & Data Analysis. 2016; 103: 316-329.

27. Bellio R, Guolo A. Integrated likelihood inference in small sample meta-analysis for continuous outcomes. Scandinavian Journal of Statistics. 2016; 43: 191-201.

28. Chatterjee K, Mukherjee D. An improved integrated likelihood population size estimation in dual-record system. Statistics & Probability Letters. 2016; 110: 146-154.es