Università
degli Studi
di Padova

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXIV

INTEGRATED LIKELIHOOD FOR THE TREATMENT OF NUISANCE PARAMETERS

Direttore della Scuola: Prof.ssa Alessandra Salvan

Supervisore: Prof. Nicola Sartori

Co-supervisori: Prof.ssa Alessandra Salvan, Prof. Thomas A. Severini

Dottorando: Riccardo De Bin

31 January 2012

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXIV

INTEGRATED LIKELIHOOD FOR THE TREATMENT OF NUISANCE PARAMETERS

Direttore della Scuola: Prof.ssa Alessandra Salvan

Supervisore: Prof. Nicola Sartori

Co-supervisori: Prof.ssa Alessandra Salvan, Prof. Thomas A. Severini

Dottorando: Riccardo De Bin

31 January 2012

Acknowledgements

First of all, I would like to thank my supervisor, Prof. Nicola Sartori, and my co-supervisors, Prof. Alessandra Salvan and Prof. Thomas A. Severini, for all the help and all the suggestions provided during the PhD program and the development of this thesis (including the final rush...). In particular, I greatly thank Prof. Thomas A. Severini for the possibility to visit the Department of Statistics at Northwestern University. About my period in Evanston, I would like to thank also Joe, for his essential help, and Arend, for all the time spent together and the endless discussions about statistics, politics and beer. Thanks also to JingSi, for her support during one of the toughest periods of my life (and, obvously, for the Chinese swearwords) and to all the other guys of the Stats Department. Last but not least (quite the contrary), a special thank to Vanessa.

I do not even start to list all the people whom I should thank in Italy, the space would not be enough. Let me just mention Bustra and the other guys of the XXIV cycle, A+, Maestro, Marlies, Monjed and Tony: my cycle is different!

Abstract

In literature, several tools have been proposed to make inference about a parameter of interest ψ in presence of nuisance parameters. Among these, the integrated likelihood seems to gain popularity. Commonly used in Bayesian inference, the integrated likelihood has been recently studied also under the frequentist paradigm. We contribute to this analysis studying first its properties in presence of many nuisance parameters, and in particular in situations when the number of nuisance parameters increases with the sample size. In this setting, indeed, the usual inferential tools, based on the profile likelihood, may perform poorly, and the use of the integrated likelihood can be an alternative to higher order methods. In particular, we focus on the asymptotic behaviour of the signed square root integrated likelihood ratio statistic, studied in a two-index asymptotics setting, in which both the sample size and the dimension of the nuisance parameter increase to infinity. As a second topic of the thesis, when there is a sufficient statistic for the nuisance parameter, we study conditions of equivalence of integrated and conditional likelihoods. Finally, some efforts are done to study the effect of the presence of nuisance parameters on pairwise likelihood and on the related score function. A correction useful to reduce the profile pairwise score bias is presented.

Sommario

In letteratura, vari strumenti sono stati introdotti per fare inferenza su un parametro di interesse ψ in presenza di parametri di disturbo. Tra questi, la verosimiglianza integrata sembra guadagnare popolarità. Usata comunemente nell'inferenza bayesiana, la verosimiglianza integrata è stata recentemente oggetto di studi approfonditi anche in ambito frequentista. Il contributo della tesi in questo ambito consiste in primo luogo nello studiare le proprietà della verosimiglianza integrata in presenza di parametri di disturbo con dimensione elevata, in particolare in situazioni in cui il numero dei parametri di disturbo cresce all'aumentare della numerosità campionaria. In questo contesto, infatti, gli strumenti inferenziali usuali, basati sulla verosimiglianza profilo, possono fornire risultati inaccurati, e l'uso della verosimiglianza integrata risulta una valida alternativa a strumenti basati su approssimazioni asintotiche di ordine più elevato. Particolare attenzione è rivolta all'analisi del comportamento asintotico della statistica radice con segno del rapporto di verosmiglianza integrata, studiata in un doppio ordine asintotico, in cui sia la numerosità campionaria, sia la dimensione del parametro di disturbo, divergono. In presenza di una statistica sufficiente per il parametro di disturbo, inoltre, sono studiati i casi di equivalenza tra la verosimiglianza integrata e condizionata. Infine, sono presentati alcuni contributi relativi allo studio degli effetti della presenza di parametri di disturbo sulla verosimiglianza a coppie e sulla relativa funzione punteggio profilo, per la quale è presentata una correzione utile per ridurne la distorsione.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Elimination of nuisance parameters is a central issue in statistical inference. Many approaches have been developed to face this problem based on pseudo-likelihood functions, including marginal likelihood, conditional likelihood, profile likelihood and its modifications; see, e.g., Pace and Salvan (1997, Chapter 4) and Severini (2000, Chapters 8 and 9). Another type of pseudo-likelihood is the integrated likelihood function (Kalbfleisch and Sprott, 1970; Liseo, 1993; Berger et al., 1999; Severini, 2000, 2007, 2010), which is of the form

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

where ψ is the parameter of interest, λ is the nuisance parameter, Λ is the parameter space for λ, assumed here independent of ψ, $L(\cdot)$ is the likelihood function and $g(\lambda; \psi)$ is a weight function for λ.

While this approach is quite common in Bayesian analysis, with $g(\lambda; \psi)$ a conditional prior density of λ given ψ, it is less used under the frequentist paradigm. A notable exception is inference on the index parameter of a group family where the marginal likelihood can be represented as an integrated likelihood. A more general contribution to the study of integrated likelihood in non-Bayesian inference is Severini (2007).

The integrated likelihood has the advantage that, unlike marginal and conditional likelihoods, it is always available and, unlike pseudo-likelihoods built on profile likelihood, it is based on averaging rather than maximization. The problems related to the maximization approach are investigated, for example, in Berger et al. (1999). On the other hand, the primary drawback of the integrated likelihood is that the weight function must be chosen (Severini, 2007). The integrated likelihood is also related to mixture models (Lindsay, 1980, 1985), random effects models and empirical Bayes techniques (Robbins, 1955; Maritz, 1970; Morris, 1983).

When the number of nuisance parameters increases with the sample size difficulties in inference procedures arise, and it is well known that the usual maximum likelihood estimator for the parameter of interest could be inconsistent. Among the techniques developed to overcome this issue, known as the incidental parameter problem, the conditional likelihood has a prominent role. In many situations, the integrated likelihood shows a similar behaviour, suggesting the existence of some common ground between the two approaches (Rice, 2008). Such similarities are studied also by Lindsay et al. (1991), Neuhaus et al. (1994) and Rice (2004). In particular, Rice (2004, 2008) shows that, in the Rasch model and matched pairs case-control studies, inference based on a conditional likelihood is the same as making inference based on an integrated likelihood with a specific weight function.

Since it depends only on the parameter of interest, the integrated likelihood can be also used to construct a likelihood ratio statistic. This statistic presents some advantages with respect to the standard likelihood ratio statistic. These gains are well explained in Severini (2010). We can recall that, from a decision theory point of view, integrated likelihood functions have the same type of optimality properties which hold for the usual likelihood in models without a nuisance parameter (Wald, 1950, Section 5.1) and, since the integrated likelihood is based on averaging rather than maximization, the use of the integrated likelihood sometimes avoids some computational problems.

The nuisance parameter problem affects also the composite likelihood, as shown, for instance, in Pakel et al. (2011). This pseudo-likelihood, first studied by Besag (1974) and Lindsay (1988), is an inference function derived by multiplying a collection of component likelihoods (Varin et al., 2011), and it is defined as

$$cL(\psi, \lambda) = \prod_{j=1}^{J} L_j(\psi, \lambda)^{w_j},$$

where $L_j(\psi, \lambda)$, $j = 1, \dots, J$, is the likelihood related to a marginal or conditional event $\mathcal{A}_j$, and w_j are nonnegative weights to be chosen (Lindsay, 1988; Varin et al., 2011). One of the most used composite likelihoods is the pairwise likelihood (see, for instance, Cox and Reid, 2004), based on the marginal distribution of two-dimensional components of a multidimensional random variable.

In the absence of nuisance parameters the composite score function is unbiased. This is not true when constrained estimates are plugged-in for λ. A modification for general estimating equations is given in Severini (2002) and this idea could be applied to composite likelihoods.

1.2 Main contributions of the thesis

The principal contribution of the thesis is the study of the properties of inferential quantities based on integrated likelihoods in presence of many nuisance parameters. In Chapter 3, in particular, we focus on the behaviour of the signed square root of the integrated likelihood ratio statistic (Severini, 2010) in a two-index asymptotics framework (Barndorff-Nielsen, 1996; Sartori, 2003), underlining how the distribution of the statistic depends on the choice of the weight function, and specifying how to choose it in order to obtain good asymptotic results.

Furthermore, in Chapter 4, with reference to equivalence between conditional and integrated likelihood, we present a generalization of Rice (2004, 2008) results to an exponential family framework, underlining the role of the moment generating function and the conditional moment generating function. This allows us to extend Rice's result to the case where the sufficient statistic could be continuous, assuming that the weight function can depend on the data, as in, among others, Wasserman (2000) and Severini (2007). Finally, we show how an approximation for such a weight function leads to the modified profile likelihood (Barndorff-Nielsen, 1983).

For a discrete sufficient statistic, we also study the relationship between conditional and integrated likelihood when the weight function related to the latter depends on an hyperparameter. We indicate which conditions assure the equivalence between the two likelihoods, reconsidering the results by Lindsay et al. (1991) an Rice (2004) from a different point of view, namely as a reparameterization problem.

Finally, in Chapter 5, we study the pairwise profile score function, in order to find a correction that reduces its bias.

Chapter 2

Preliminaries and notation

2.1 Nuisance parameters and pseudo-likelihoods

Parametric statistical inference studies how to analyse data through models where the information about the phenomena is summarized through quantities called *parameters*. Usually, not all the parameters are of primary interest and some are considered only to appropriately describe variability in the population. These parameters are called *nuisance parameters*. The simplest setting is when a parametric model $\mathcal{F} = \{p_Y(y;\theta), y \in \mathcal{Y}, \theta \in \Theta\}$ is assumed, and the p-dimensional parameter θ is partitioned as $\theta = (\psi, \lambda)$, with ψ a k-dimensional parameter of interest, and λ a $(p-k)$-dimensional nuisance parameter. Hereafter we denote with $p_Y(y;\theta)$ the probability density of the random variable Y, with $\mathcal{Y}$ the sample space and with Θ the parameter space.

We will also focus on a particular type of nuisance parameters, namely the so called *incidental parameters* (Neyman and Scott, 1948). Consider data $y = (y_1, \dots, y_n)$, where y_i, $i = 1, \dots, n$, is a realization of a random variable Y_i with density $p_{Y_i}(y_i;\psi,\lambda_i)$. While ψ is common to all observations, and for this reason called *structural parameter*, the nuisance component λ_i is specific to each Y_i, and leads to a nuisance parameter of the form $\lambda = (\lambda_1, \dots, \lambda_n)$. In particular, this means that the dimension of λ depends on the sample size n.

Unfortunately, the presence of nuisance parameters often affects inferential procedures about the parameter of interest, especially in the incidental parameters case. In order to deal with this issue, several solutions have been proposed. Here we focus on those based on the notion of *pseudo-likelihood*. In broad generality, with pseudo-likelihood we indicate any function of the data and the parameter which behaves, at least in some respects, as if it were a genuine likelihood. In Section 2.4 we will discuss a particular kind of pseudo-likelihood which is called *composite likelihood*. When dealing with nuisance parameters, instead, we typically refer to pseudo-likelihoods which depend on the data and the parameter of interest only. Obviously, the attractiveness of basing inference on a pseudo-likelihood increases with the complexity of the structure of the nuisance component (Pace and Salvan, 1997, Section 4.3).

A first way to construct a pseudo-likelihood for a parameter of interest is based on a statistical model defined as a reduction of the original model

$\mathcal{F}$ and it is related to the notion of partial sufficiency and partial ancillarity (Barndorff-Nielsen and Cox, 1994, Sections 2.3 and 2.5).

Suppose that y is a one-to-one function of (s, t), such that

$$p_{T,S}(t, s; \psi, \lambda) = p_{T|S=s}(t; s, \psi, \lambda) p_S(s; \psi, \lambda). \tag{2.1}$$

If s is partially sufficient for λ, then

$$p_{T|S=s}(t; \psi, \lambda, s) = p_{T|S=s}(t; \psi, s), \tag{2.2}$$

and we can construct a likelihood based on it, obtaining the *conditional likelihood*,

$$L_C(\psi) = p_{T|S=s}(t; \psi, s).$$

A somehow complementary instance is when s is partially distribution constant for λ, then

$$p_S(s; \psi, \lambda) = p_S(s; \psi),$$

and inference can be based on the *marginal likelihood*,

$$L_M(\psi) = p_S(s; \psi).$$

Both conditional and marginal likelihoods are genuine likelihoods, so they satisfy all the standard properties of a likelihood. Unfortunately, they are not generally available outside special families of distributions. Furthermore, besides the special situation

$$p_{T,S}(t, s; \psi, \lambda) = p_{T|S=s}(t|s; \psi) p_S(s; \lambda),$$

when the likelihood has separable parameters and $L(\theta) = L_1(\psi; t|s) L_2(\lambda; s)$, there could be a loss of information due to the discarded factor (see, for instance, Severini, 2000, Section 8).

A different, and more general, idea to construct a pseudo-likelihood for ψ is to substitute the nuisance parameter with a consistent estimate in the original likelihood. The *profile likelihood*,

$$L_P(\psi) = L(\psi, \hat{\lambda}_\psi),$$

where $\hat{\lambda}_\psi$ is the maximum likelihood estimate of λ for fixed ψ, is the most notable case. Although the profile likelihood has good properties, it is not a genuine likelihood, since it is not deduced from a density function, and this may have some drawbacks. In particular, the profile score, namely the first derivative of the logarithm of the profile likelihood, is biased, and this leads to possible inconsistency of the maximum likelihood estimator for ψ, when dealing with incidental nuisance parameters (Neyman and Scott, 1948).

Several modifications of the profile likelihood have been proposed to handle this issue. The idea is to correct the profile likelihood in order to take into account the lack of information due to the presence of nuisance parameters, and to penalize the profile likelihood in order to reduce its score bias. We recall, among the others, the modifications proposed by Barndorff-Nielsen (1983, 1994, 1995), Cox and Reid (1987) and McCullagh and Tibshirani (1990). In particular, the *approximate conditional likelihood* by Cox and Reid (1987) stands out due to its simple form, since it involves only the information matrix for the nuisance parameter,

$$L_{AC}(\psi) = L(\psi, \hat{\lambda}_\psi)| - l_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{-1/2}. \tag{2.3}$$

Here $-l_{\lambda\lambda}(\psi, \lambda) = -\partial^2 \log L(\psi, \lambda)/\partial\lambda\partial\lambda^T$ denotes the block of the information matrix related to the nuisance parameter.

Unfortunately, $L_{AC}(\psi)$ depends on the nuisance parameterization and requires ψ and λ to be orthogonal (see, for instance, Barndorff-Nielsen and Cox, 1994, Section 2.7). In this case, moreover, it approximates the *modified profile likelihood* (Barndorff-Nielsen, 1983). The latter, based on the p^*-formula (Barndorff-Nielsen, 1980, 1983), requires the specification of an ancillary statistic (Barndorff-Nielsen and Cox, 1994, Section 2.5). In this case *ancillary* means not only distribution constant, but it is required that the statistic, together with the maximum likelihood estimator, constitutes a sufficient statistic. Writing the likelihood in the form

$$L(\psi, \lambda; y) = L(\psi, \lambda; \hat{\psi}, \hat{\lambda}, a),$$

where a is the ancillary statistic and $(\hat{\psi}, \hat{\lambda})$ the maximum likelihood estimate, the modified profile likelihood is

$$L_{MP}(\psi) = L(\psi, \hat{\lambda}_\psi) \frac{| - l_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{1/2}}{|l_{\lambda;\hat{\lambda}}(\psi, \hat{\lambda}_\psi)|},$$

where $l_{\lambda;\hat{\lambda}}(\psi, \lambda) = \partial^2 \log L(\psi, \lambda; \hat{\psi}, \hat{\lambda}, a)/(\partial\lambda\partial\hat{\lambda}^\top)$ is a sample space derivative (see also Severini, 2000, Section 6.2).

The modified profile likelihood can be obtained both as an approximation of a conditional likelihood and of a marginal likelihood for ψ, if they exist (see, for instance, Severini, 2000, Sections 9.3.2 and 9.3.3). In both cases, the order of the approximation is $O(n^{-3/2})$ in the moderate-deviation range, i.e. for ψ of the form $\psi = \hat{\psi} + O_p(n^{-1/2})$, and $O(n^{-1})$ in the large-deviation range, that is for ψ of the form $\psi = \hat{\psi} + O_p(1)$.

The main advantages of the modified profile likelihood over the approximate conditional likelihood (2.3) lie in its invariance under interest-respecting reparameterization and in the fact that an explicit orthogonal parameterization is not required. On the other hand, an ancillary statistic must be available. To avoid this point, several approximations have been proposed, for example by Barndorff-Nielsen (1994) and Severini (1998a). The order of error of these approximations is $O(n^{-1})$ (for a review, see Severini, 2000, Section 9.5).

Here with *invariance under reparameterization* we denote the property which assures that the conclusion of a statistical analysis is not changed by different formulations of the parameter, such as $\tilde{\theta} = \tilde{\theta}(\theta)$, a one-to-one function of θ. Furthermore, when $\theta = (\psi, \lambda)$, we denote with *interest-respecting reparameterization* the transformation $\tilde{\theta}(\theta) = (\tilde{\psi}, \tilde{\lambda})$, where $\tilde{\psi}$ is a one-to-one function of ψ and $\tilde{\lambda}$ is a function of both ψ and λ (see, for instance, Barndorff-Nielsen and Cox, 1994, Section 1.5).

A simpler idea to adjust the profile likelihood has been pursued by McCullagh and Tibshirani (1990). The proposed modification, in fact, requires to compute via simulation (parametric bootstrap) the first two moments of the profile score function and to use them in order to centre and rescale it. When ψ is scalar, the integral with respect to ψ of such adjusted profile score gives the *adjusted profile log-likelihood.* A first order approximation for the adjustment, which does not required simulation, is also provided.

Finally, an alternative proposal is the *generalized profile likelihood* (Severini, 1998b), that consists in substituting the constrained maximum likelihood estimate $\hat{\lambda}_\psi$ with an alternative $\tilde{\lambda}_\psi$, that may be in some sense superior, such as the empirical Bayes estimate when the dimension of λ is large (see, e.g., Morris, 1983).

Summing up, the main idea in profile likelihood is to eliminate the effect of nuisance parameter through a maximization step, the computation of the constrained maximum likelihood estimate. The various modifications presented, moreover, try to adjust the profile likelihood taking into account the nuisance parameter estimate uncertainty. In practice, they penalize the values of ψ for which the information about λ is relatively large.

A different way to eliminate the nuisance parameters is to exploit averaging instead of maximization. The related pseudo-likelihood is the *integrated likelihood*. It is based on the idea that we can summarize the set of likelihoods $\{L(\theta) : \theta \in \Theta(\psi)\}$, where $\Theta(\psi)$ is the subset of the parameter space such that $\psi = \psi(\theta)$, by its average value with respect to some weight function $g(\lambda; \psi)$ over $\Theta(\psi)$ (Severini, 2007). If $\Theta(\psi) = \{(\psi, \lambda) : \lambda \in \Lambda\}$, then

$$L_I(\psi) = \int_\Lambda p_Y(y; \psi, \lambda) g(\lambda; \psi)\, d\lambda. \tag{2.4}$$

Thanks to this averaging step, the integrated likelihood automatically incorporates nuisance parameter uncertainty (Berger et al., 1999).

Originally developed in the Bayesian framework, the integrated likelihood has been studied in the non-Bayesian field primarily by Severini (2007). In that paper, the author shows that, when the nuisance parameter is strongly unrelated to the parameter of interest and the weight function does not depend on ψ, the integrated likelihood preserves several desirable properties of a genuine likelihood, such as Bartlett's identities, invariance and insensitivity to the choice of weight function. If we recall that it is always available, unlike, for example, marginal and conditional likelihood, and that averaging permits to avoid some possible problems due to the maximization step of other pseudo-likelihoods (see, for instance, Berger et al., 1999, Section 2.1), we can state that the integrated likelihood could be a useful tool in order to make inference about the parameter of interest.

Severini (2007) provided also a special reparameterization, called *zero-score expectation* parameterization, which allows to easily obtain a nuisance parameter strongly unrelated to ψ. The new nuisance parameter $\phi = \phi(\psi, \lambda; \hat{\psi})$ is the solution of the equation

$$E_{(\psi_0, \lambda_0)}[l_\lambda(\psi, \lambda)]\big|_{(\psi_0, \lambda_0) = (\hat{\psi}, \phi)} = 0, \tag{2.5}$$

where $l_\lambda(\psi, \lambda)$ is the score function component related to λ and the subscript in the operator E means that the expected value is taken when the true parameter value is (ψ_0, λ_0). The main feature of ϕ is that it depends on the data. While this idea can cause some problems for Bayesian inference, it can be used safely in the non-Bayesian framework (Severini, 2007).

The integrated likelihood is also related with random effects models, mixture models and empirical Bayesian techniques, especially when the weight function depends on further parameters, called *hyperparameters*.

2.2 Integrated likelihood and signed square root likelihood ratio statistic

As a pseudo-likelihood, the integrated likelihood depends only on the parameter of interest and on the data, so it can be used in the construction of the usual inferential tools. In particular, for a scalar ψ, we will focus on the *signed square root integrated likelihood ratio statistic*,

$$\bar{R} = \text{sgn}(\bar{\psi} - \psi)\sqrt{2[\log L_I(\bar{\psi}) - \log L_I(\psi)]}, \tag{2.6}$$

where $\bar{\psi}$ is the maximizer of $L_I(\psi)$.

Asymptotic properties of $\bar{R}$ have been studied by Severini (2010). In that paper, the asymptotic standard normal distribution of the statistic is shown, and a correction is provided, through the approach developed by Sweeting (1995, 1996). This is based on a term

$$Q = \frac{g(\hat{\lambda}; \psi)}{g(\hat{\lambda}; \hat{\psi})} \frac{| - l_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})|^{1/2}}{| - l_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{1/2}},$$

which assures the second order normality of the modified signed square root integrated likelihood ratio statistic. Recalling the quantities

$$u_I = \frac{\left| \frac{\partial}{\partial \hat{\psi}} \{ l_P(\hat{\psi}) - l_P(\psi) \} \right|}{| - l_{P\psi\psi}(\hat{\psi})|^{\frac{1}{2}}},$$

where $l_P(\psi) = \log L_P(\psi)$ and $-l_{P\psi\psi}(\psi) = -\frac{\partial^2}{\partial \psi^2} l_P(\psi)$, and

$$u_P = \frac{|l_{\lambda;\hat{\lambda}}(\psi, \hat{\lambda}_\psi)|}{| - l_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{\frac{1}{2}} | - l_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})|^{\frac{1}{2}}},$$

derived by Pierce and Peters (1992), which represent the deviation from normality due to a small amount of information in the sample and the effect of nuisance parameters, respectively, the author defines the modified integrated likelihood ratio statistic as

$$\bar{R}^{*} = \bar{R} + \bar{R}^{-1} \log \left(\frac{u_P \, Q \, u_I}{R} \right).$$

Severini (2010) shows that $\bar{R}^*$ agrees to the second order with the modified directed likelihood ratio statistic R^* (Barndorff-Nielsen, 1986). Since R^* is know to be asymptotically normally distributed with error $O_p(n^{-3/2})$ in the moderate-deviation case and $O_p(n^{-1})$ in the large-deviation case, second order asymptotic normality of $\bar{R}^*$ follows. Severini (2010) shows, however, that second order asymptotic normality can be achieved also for the regular integrated likelihood ratio statistic without the need of any correction, through a wise choice of weight function. In the orthogonal parameter case, indeed, it is possible to drop out the bias due to the presence of nuisance parameter forcing the bias due to the choice of the weight function to be its opposite. The desired weight function is such that

$$\frac{\partial}{\partial \psi} \log \, g(\lambda; \psi) = \frac{1}{6} \frac{E_{(\psi,\lambda)}[l_{\psi}(\psi, \lambda)^3]}{E_{(\psi,\lambda)}[l_{\psi}(\psi, \lambda)^2]}, \tag{2.7}$$

where $l_{\psi}(\psi, \lambda) = \partial \log l(\psi, \lambda)/\partial \psi$.

These properties of $\bar{R}$ have been studied only in the ordinary asymptotic framework. When there are stratum nuisance parameters, however, the usual tools are not sufficient to understand the behaviour of $\bar{R}$ and it is necessary to study it in the so called *two-index asymptotics* (Barndorff-Nielsen, 1996), i.e. an asymptotic scenario which allows both the within-stratum sample size (m) and the number of strata (q), and consequently the dimension of nuisance parameter, to diverge.

The study of asymptotic properties for stratified data in two-index asymptotics setting has been developed for likelihood statistics based on the modified profile likelihoods by Sartori (2003). In particular, it is shown that the likelihood ratio, the score and the Wald statistics based on a modified profile likelihood have, in a two-index asymptotics framework, a χ^2 distribution, if $\frac{q}{m^3} = o(1)$. Differently, ordinary statistics based on profile likelihood are guaranteed to have that distribution only if $\frac{q}{m} = o(1)$. This means that there is a range of cases where likelihood statistics based on modified profile likelihoods perform well, while the ones based on profile likelihood do not. Results about profile likelihood based statistics confirm and extend outside the regular exponential family case the analysis of Portnoy (1988).

Focusing on the signed square root likelihood ratio statistic, we will see that similar results hold for the integrated likelihood based statistics. In the two-index asymptotics setting, moreover, Sartori et al. (1999) show that R_{MP}, the likelihood ratio statistic based on L_{MP}, achieves most of the gain attained by R^*. Starting from decomposition of R^*

$$R^* = R + R^{-1} \log \frac{u_I}{R} + R^{-1} \log u_P$$

proved in exponential families by Pierce and Peters (1992) and in general by Barndorff-Nielsen and Cox (1994, Section 6.6.4), Sartori et al. (1999) note that R_{MP} is equal, to second asymptotic order, to the part of R^* including the correction for the effect of nuisance parameter ($R^{-1} \log u_P$). Since in the two-index asymptotics setting the remaining part involving u_I is $O_p(1/\sqrt{mq})$, this means that the distributions of R_{MP} and of R^* become close to each other as both the sample size or the number of strata increases, despite the fact that the dimension of the nuisance parameter increases with the sample size.

2.3 Integrated versus marginal and conditional likelihood

The conditional and the marginal likelihoods are genuine likelihoods, but they require the availability of a partially sufficient and partially distribution constant statistics, respectively. In practice, the use of this kind of pseudo-likelihoods is restricted to multiparameter exponential and to composite group families (Pace and Salvan, 1997, Section 5.4 and Section 7.5). Moreover, even when model reduction is possible, it may be computationally difficult to actually obtain such likelihoods. It would be interesting to understand whether marginal and conditional likelihoods could be achieved through an integrated likelihood.

It is well known that this is possible for marginal likelihood. As stated, for instance, in Barndorff-Nielsen and Cox (1994, Section 2.8), the marginal likelihood agrees with an integrated likelihood where the weight function is chosen to be the *right invariant prior distribution.* In a composite group family, the marginal likelihood for the shape parameter ψ, based on the maximal invariant, can be always represented as an integrated likelihood with respect to the right invariant measure (Pace and Salvan, 1997, Section 7.7).

The question is unresolved for the conditional likelihood. The literature related to this topic, indeed, focuses primarily on the equivalence between the conditional and the integrated maximum likelihood estimators. This equivalence is proved, for example, by Lindsay et al. (1991) for the Rasch model, provided that a condition called *PD concordance* holds. A similar result is shown by Neuhaus et al. (1994) for binary matched pairs data.

Some cue of the equivalence between the conditional and integrated approach can be instead found in Andersen and Madsen (1977) and Andersen (1980, Chapter 6), at least with reference to the Rasch model. However, the authors were again primarily interested in the estimation of the parameters, and the equivalence between the entire functions was not delved into. This aspect was investigated by Rice (2004) in the Rasch model, and by Rice (2008) for matched pairs case-control studies. In particular, the author considers the distribution of the sufficient statistic for the nuisance parameter, obtaining some marginal probabilities that depend both on ψ and λ. Since λ is eliminated integrating with respect to a weight function, he points out that it is sufficient to adapt the weight function in order to drop out the dependence of marginal probabilities on ψ.

Severini (1999) compares the pseudo-likelihoods hitherto considered with Bayesian elimination of nuisance parameter. In particular, for multiparameter exponential families, a prior which assures third order asymptotic equivalence between conditional and integrated likelihoods is obtained. This is the determinant of the Hessian with respect to λ of the cumulant generating function. Finally, in the case of orthogonal nuisance parameter, it is shown that the profile likelihood is second order asymptotically equivalent to an integrated likelihood with a prior built multiplying a specific factor to an arbitrary function of λ. As noted firstly by (Sweeting, 1987), moreover, the approximate conditional likelihood is third order asymptotically equivalent to an integrated likelihood with respect to a constant prior for λ which should be independent of ψ.

2.4 Composite likelihood and nuisance parameters

A different kind of pseudo-likelihood, not related to the presence of nuisance parameters, is the composite likelihood. This pseudo-likelihood has been introduced in order to deal with models where complex interdependencies are involved (for a review, see Varin et al., 2011). Although the first instance was proposed by Besag (1974) in spatial statistics, the term *composite likelihood* was successfully introduced by Lindsay (1988), with the purpose of better describing the method of its construction. Indeed, denoting by $\{\mathcal{A}_1, \ldots, \mathcal{A}_J\}$ a set of marginal or conditional events with associated likelihoods $L_j(\theta; y)$, $j = 1, \ldots, J$, where y is a realization of the random variable Y with density $p_Y(y; \theta)$, the composite likelihood is constructed as a weighted product of J components,

$$cL(\theta) = \prod_{j=1}^{J} L_j(\theta)^{w_j}.$$

Here, w_j are nonnegative weights to be chosen (Lindsay, 1988; Varin et al., 2011).

As for the standard likelihood, we can compute the composite log-likelihood as the logarithm of the composite likelihood, $cl(\theta) = \log cL(\theta)$, and its maximum, called the maximum composite likelihood estimate. In standard problems, this is the solution of the composite score function equal to zero, which is a linear combination of the scores associated with each log-likelihood term $l_j(\theta) = \log L_j(\theta)$ (Varin et al., 2011).

Since the likelihood terms forming the pseudo-likelihood are not independent and, anyway, due to the fact that the resulting function is not proportional to a density function, we can see the composite likelihood as a misspecified likelihood. Denoting with $cl_\theta(\theta)$ and $-cl_{\theta\theta}(\theta)$ the composite version of the score function and the observed information, respectively, we see that the sensitivity matrix

$$H(\theta) = E_\theta[-cl_{\theta\theta}(\theta)]$$

differs from the variability matrix

$$J(\theta) = E_\theta[cl_\theta(\theta) cl_\theta(\theta)^T],$$

or, in other words, that the second Bartlett identity does not hold. On the other hand, since all the score components are unbiased, the first Bartlett identity is still valid. This, under generally mild conditions, guarantees the consistency of the composite likelihood estimator.

Furthermore, the failure of the second Bartlett identity has some consequence on the asymptotic theory. For example, given n independent and identically distributed observations from $Y_i \in R^d$, $i = 1, \dots, n$, with density $p_{Y_i}(y_i; \theta)$, as $n \to \infty$ and d fixed, the asymptotic distribution of the composite maximum likelihood estimator is

$$\sqrt{n}(\hat{\theta}^c - \theta) \xrightarrow{d} N_p(0, (H(\theta)J^{-1}(\theta)H(\theta))^{-1}).$$

The composite version of both Wald and score statistics have the usual asymptotic χ^2_k distribution (Molenberghs and Verbeke, 2005, Section 9.3). Nevertheless, they suffer from the same practical limitations of their ordinary version: the former is not invariant under reparameterization, while the latter may be numerically unstable (Varin et al., 2011). On the other hand, the composite likelihood ratio statistic, which is typically preferable, has the non-standard asymptotic distribution

$$\sum_{j=1}^{k} \lambda_j Z_j^2,$$

where $Z_1, \dots, Z_k$ are independent normal random variables and $\lambda_1, \dots, \lambda_k$ are the eigenvalues of a matrix related to $H(\theta)$ and $J(\theta)$ (Foutz and Srivastava, 1977, 1978; Kent, 1982). This makes it hard to deal with. For this reason, several adjustments have been proposed for the likelihood ratio statistic, based on first (Rotnitzky and Jewell (1990); Molenberghs and Verbeke (2005, Section 9.3.3)), first and second (Satterthwaite, 1946; Varin, 2008) or higher (Wood, 1989; Lindsay et al., 2000) order moment matching. Finally, in order to recover the usual asymptotic χ^2_k distribution, vertical rescaling has been exploited by Chandler and Bate (2007) and Pace et al. (2011).

In this thesis we will focus our attention on what happens in the presence of nuisance parameters. As for the ordinary likelihood, when $\theta = (\psi, \lambda)$ has a component ψ of interest, we can consider the profile version of the composite likelihood,

$$cL_P(\psi) = cL(\psi, \hat{\lambda}^c_\psi),$$

with $\hat{\lambda}^c_\psi$ denoting the composite constrained estimate for λ given ψ. In general, the first Bartlett identity does not hold anymore, i.e. the corresponding score function is biased,

$$E_{(\psi,\lambda)}[\frac{\partial}{\partial\psi}\log cL_P(\psi)] \neq 0.$$

This may cause some problems, especially when the dimension of the nuisance parameter is large (see, e.g., Pakel et al., 2011). The dependence structure among the components of the composite likelihood, however, does not allow us to use the ordinary score bias corrections.

In Chapter 5, we will consider the score bias of the profile pairwise likelihood. The pairwise likelihood (Cox and Reid, 2004) is a special instance of composite likelihood, constructed using as likelihood components the likelihoods based on the two-dimensional marginal distribution of all pairs (Y_r, Y_s), $r = 1, \dots, d-1$, $s = r, \dots, d$, of a random vector $Y = (Y_1, \dots, Y_d)$. Denoting by $p_{Y_rY_s}(y_r, y_s; \psi, \lambda)$ the two-dimensional density, the pairwise likelihood is defined as

$$pL(\psi, \lambda) = \prod_{r=1}^{d-1}\prod_{s=r+1}^{d} p_{Y_rY_s}(y_r, y_s; \psi, \lambda)^{w_{rs}}. \qquad (2.8)$$

Obviously, it has all the properties and issues of a general composite likelihood just described. Nevertheless, it is worth studying the pairwise likelihood because it combines the simplicity of its formulation (it involves only two-dimensional density functions) with the possibility to make inference on parameters related to dependence.

Chapter 3

Integrated likelihood ratio statistic in models with stratum nuisance parameters

3.1 Introduction

Inference about ψ proceeds by treating the integrated likelihood as a genuine likelihood function in order to form point estimates, confidence intervals and so on. In this chapter, we focus on the properties of the integrated signed root likelihood ratio statistic introduced in Section 2.2 (see formula (2.6)). In models in which the dimension of the nuisance parameter is fixed, first-order asymptotic theory shows that $\bar{R}$ is asymptotically distributed according to a standard normal distribution. In Severini (2010) it is shown that $\bar{R}$ may be modified so that the resulting statistic is asymptotically standard normal to a higher order of approximation.

Here we study how integrated likelihood ratio statistic performs in stratified models, in particular when the number of strata is large relative to the total sample size. To this end, we will consider a two-index asymptotics setting. First, in Section 3.2, we introduce the framework where we evaluate inferential procedures. Section 3.3 shows the Laplace approximations that we use in the following sections. In Section 3.4 we analyse the integrated score function, while the signed square root likelihood ratio statistic is studied in Section 3.5. Some ideas about the choice of the weight function are provided in Section 3.6. Section 3.7 contains several examples, while some conclusions are drawn in Section 3.8.

3.2 Asymptotic framework

Let us consider a stratified model with Y_{ij} independent random variables having density $p_{ij}(y_{ij};\psi,\lambda_i)$, $i=1,\dots,q$, $j=1,\dots,m$. In the asymptotic scenario in which both m and q approach infinity, the nuisance parameter of the model is the infinite sequence $\lambda_1,\lambda_2,\dots$. It follows that any asymptotic properties, such as consistency or asymptotic normality, could depend on the properties of this sequence (Portnoy, 1988). For instance, consider an exponential family for $Y_{ij}=(X_{ij},Z_{ij})$, with ψ and λ scalar parameters, giving log-likelihood contribution $x_{ij}\psi+z_{ij}\lambda_i-K(\psi,\lambda_i)$, where K is the cumulant generating function of the exponential family. Then the score function for ψ is given by $l_\psi=m\sum_{i=1}^{q}\{\bar{x}_i-K_\psi(\psi,\lambda_i)\}$ where $\bar{x}_i$ is the sample mean of the x-values in the i-th stratum and K_ψ is the derivative of

K with respect to ψ. Then $l_\psi/\sqrt{mq}$ has cumulant generating function

$$\sum_{i=1}^{q} m\{K(\psi + t/\sqrt{mq}, \lambda_i) - K(\psi, \lambda_i) - t\sqrt{\frac{m}{q}}K_\psi(\psi, \lambda_i)\}$$
$$= \frac{1}{q}\sum_{i=1}^{q} K_{\psi\psi}(\psi, \lambda_i)\frac{t^2}{2} + \frac{1}{q^{\frac{3}{2}}}\sum_{i=1}^{q} K_{\psi\psi\psi}(\psi, \lambda_i)\frac{t^3}{6} + \cdots .$$

It follows that asymptotic normality of l_ψ requires conditions on $\lambda_1, \lambda_2, \ldots$ through conditions on the sequences $K_{\psi\psi}(\psi, \lambda_i)$, $K_{\psi\psi\psi}(\psi, \lambda_i)$, and so on.

Although there is empirical evidence regarding the values of λ_i for those strata actually observed, those λ_i are irrelevant for questions of convergence. Thus, it is reasonable to proceed as if the values of λ_i corresponding to unobserved strata have properties similar to those of the λ_i corresponding to observed strata. One way to formalize this idea is to assume that $\lambda_1, \lambda_2, \ldots$ is a sequence of realized random variables. Thus, we assume that $\lambda_1, \lambda_2, \ldots$ are independent, identically distributed random variables each with an (unknown) density function $\pi(\cdot; \psi)$; note that, in order to maintain appropriate parameterization invariance, the density function of the nuisance parameters must be allowed to depend on ψ. When we apply an interest-respecting reparameterization, indeed, the new nuisance parameter could depend on ψ, and so the Jacobian of the transformation, which should be included in $\pi(\cdot; \psi)$, depends on the parameter of interest. Furthermore, in contrast to the weight function g used to form an integrated likelihood, we require that π be a genuine density function. This type of technical device is commonly used in models with an increasing number of nuisance parameters; see, e.g., Kiefer and Wolfowitz (1956), Pfanzagl and Wefelmeyer (1982), Follmann (1988), Skovgaard (1989), and Strasser (1996). Pfanzagl (1993) gives a detailed discussion of the differences in the asymptotic theory under models with fixed and random nuisance parameters.

In the present chapter, we will use the model parameterized by $(\psi, \lambda_1, \lambda_2, \ldots)$ in developing inferential procedures for ψ. Without loss of generality, we will consider one-dimensional incidental parameters λ_i. The asymptotic properties of integrated likelihood procedures will be evaluated under the marginal model with density

$$p^{\ddagger}(y; \psi) = \prod_{i=1}^{q} \int_\Lambda p_i(y_i; \psi, \lambda)\pi(\lambda; \psi)d\lambda,$$

where $p_i(y;\psi,\lambda)$ is the joint density of $Y_i = (Y_{i1},\dots,Y_{im})$. Hence, under the marginal model, $Y_1,\dots,Y_q$ are independent random vectors with joint density $p^{\ddagger}(y;\psi)$. Since the density π is unknown, an asymptotic distribution depending on a specific value for π will not be useful for statistical inference. Thus, our goal is to consider those asymptotic properties that hold for a wide class of π. For instance, if a given statistic T is asymptotically standard normal under the marginal model, for any density π, it seems reasonable to approximate the distribution of T by a standard normal distribution.

3.3 Laplace approximation

Exact integration of the likelihood function will be possible only in exceptional cases; hence, we will rely on the use of Laplace approximations in deriving the properties of procedures based on the integrated likelihood. Let $l_I(\psi) = \log L_I(\psi)$. Since each λ_i applies only in a single stratum, a Laplace approximation for L_I can be obtained by using a Laplace approximation in each stratum and combining the results:

$$l_I(\psi) = l_P(\psi) + \sum_{i=1}^{q} \log g(\hat\lambda_{i\psi};\psi) - \sum_{i=1}^{q} \log |j_{\lambda_i\lambda_i}(\psi,\hat\lambda_{i\psi})|^{1/2} + \sum_{i=1}^{q} Q_i(\psi,\hat\lambda_{i\psi}). \tag{3.1}$$

Here $\hat\lambda_{i\psi}$ denotes the maximum likelihood estimator of λ_i for fixed ψ, $l_P(\psi)$ is the profile log-likelihood function, given by $\sum_{i=1}^{q} \log p_i(y_i;\psi,\hat\lambda_{i\psi})$, and $Q_i(\psi)$ is the $O(m^{-1})$ term in the Laplace approximation in stratum i, given by

$$\begin{aligned} Q_i(\psi) =& \log\left(1 - \frac{5}{24}\frac{|\tilde l_{\lambda_i\lambda_i\lambda_i}|^2}{|\tilde l_{\lambda_i\lambda_i}|^3} + \frac{1}{8}\frac{|\tilde l_{\lambda_i\lambda_i\lambda_i\lambda_i}|}{|\tilde l_{\lambda_i\lambda_i}|^2} + \frac{1}{2}\frac{|\tilde l_{\lambda_i\lambda_i\lambda_i}|}{|\tilde l_{\lambda_i\lambda_i}|^2}\frac{\tilde g_{\lambda_i}}{\tilde g} - \frac{1}{2}\frac{1}{|\tilde l_{\lambda_i\lambda_i}|}\frac{\tilde g_{\lambda_i\lambda_i}}{\tilde g}\right) \\ & + O(m^{-2}) \end{aligned} \tag{3.2}$$

with

$$\tilde l_{\lambda_i\lambda_i} = \left.\frac{\partial^2 l(\psi,\lambda_i)}{\partial\lambda_i^2}\right|_{\lambda_i=\hat\lambda_{i\psi}}, \quad \tilde l_{\lambda_i\lambda_i\lambda_i} = \left.\frac{\partial^3 l(\psi,\lambda_i)}{\partial\lambda_i^3}\right|_{\lambda_i=\hat\lambda_{i\psi}},$$

and so on,

$$\tilde g_{\lambda_i} = \left.\frac{\partial g(\lambda;\psi)}{\partial\lambda_i}\right|_{\lambda_i=\hat\lambda_{i\psi}}$$

and

$$\tilde{g}_{\lambda_i\lambda_i} = \left.\frac{\partial^2 g(\lambda;\psi)}{\partial\lambda_i^2}\right|_{\lambda_i=\hat{\lambda}_{i\psi}}.$$

The same type of approximation holds for the log-likelihood based on the marginal model. Let $l^\ddagger(\psi) = \log p^\ddagger(y;\psi)$. Then

$$l^\ddagger(\psi) = l_P(\psi) + \log\prod_{i=1}^q \pi(\hat{\lambda}_{i\psi};\psi) - \log\prod_{i=1}^q |j_{\lambda_i\lambda_i}(\psi,\hat{\lambda}_{i\psi})|^{1/2} + \sum_{i=1}^q V_i(\psi), \quad (3.3)$$

where $V_i(\psi)$ has the same form as $Q_i(\psi)$, with π replacing g. By combining Laplace approximations (3.1) and (3.3), we have that

$$l_I(\psi) = l^\ddagger(\psi) + \sum_{i=1}^q \log\frac{g(\hat{\lambda}_{i\psi};\psi)}{\pi(\hat{\lambda}_{i\psi};\psi)} + O\left(\frac{q}{m}\right);$$

hence, the score functions satisfy

$$l_{I\psi}(\psi) = l^\ddagger_\psi(\psi) + \sum_{i=1}^q \frac{\partial}{\partial\psi}\log\frac{g(\hat{\lambda}_{i\psi};\psi)}{\pi(\hat{\lambda}_{i\psi};\psi)} + O\left(\frac{q}{m}\right). \quad (3.4)$$

3.4 Score functions

We now consider the relationship between the score functions $l_{I\psi}(\psi)$ and $l^\ddagger_\psi(\psi)$. For $i = 1, 2, \ldots, q$, let

$$D_i(\psi) = \frac{\partial}{\partial\psi}\log\frac{g(\hat{\lambda}_{i\psi};\psi)}{\pi(\hat{\lambda}_{i\psi};\psi)}.$$

Recalling (3.4), the score functions based on l_I and $l^\ddagger$ satisfy

$$l_{I\psi}(\psi) = l^\ddagger_\psi(\psi) + \sum_{i=1}^q D_i(\psi) + O_p\left(\frac{q}{m}\right). \quad (3.5)$$

The properties of $D_i(\psi)$ depend on the properties of $g(\lambda_i;\psi)$ and $\pi(\lambda_i;\psi)$. In general, $D_i(\psi) = O_p(1)$ and has mean of order $O(1)$ so that

$$l_{I\psi}(\psi) = l^\ddagger_\psi(\psi) + O(q) + O_p\left(\frac{q}{m}\right),$$

and the score bias of $l_{I\psi}(\psi)$ under the marginal model is of order $O(q)$; this is the same order as the score bias of the profile likelihood (Sartori, 2003).

There are two cases in which the order of $\sum_{i=1}^{q} D_i(\psi)$ is smaller than $O_p(q)$. The first is when the ratio $g(\lambda_i;\psi)/\pi(\lambda_i;\psi)$ is orthogonal to ψ, i.e. the ratio does not depend on ψ, for each $i = 1, 2, \ldots, q$. Then

$$\frac{\partial}{\partial\psi} \log \frac{g(\hat{\lambda}_{i\psi};\psi)}{\pi(\hat{\lambda}_{i\psi};\psi)}$$

is of order $O_p(m^{-1/2})$ and has expected value of order $O(m^{-1})$. Since $\sum_{i=1}^{q} \sqrt{m}\{D_i(\psi) - E[D_i(\psi)]\} = O_p(\sqrt{q})$, it follows that

$$l_{I\psi}(\psi) = l^{\ddagger}_{\psi}(\psi) + O_p\left(\frac{q}{\sqrt{m}}\right)$$

and l_I has score bias of order $O(\frac{q}{m})$. Thus, for a given choice of g, there is a class of density functions π such that l_I is approximately score unbiased under the marginal model corresponding to π. Note that the order of the score bias in this case is the same as the order of the score bias of the modified profile likelihood (Sartori, 2003).

For each $i = 1, 2, \ldots, q$, let ϕ_i denote a function of (λ_i, ψ) such that ϕ_i is orthogonal to ψ. The second case occurs when the weight function for ϕ_i corresponding to g does not depend on ψ for each $i = 1, 2, \ldots, q$; in this case we will say that g is orthogonal to ψ. Then $D_i(\psi)$ has expected value of order $O(1/m)$ under the marginal model; this follows from the relationship between $l_I(\psi)$ and the adjusted profile likelihood (Cox and Reid, 1987), along with the fact that $l^{\ddagger}(\psi)$ is exactly score unbiased under the marginal model. It follows that $\sum_{i=1}^{q} D_i(\psi) = O_p(\sqrt{q})$ so that

$$l_{I\psi}(\psi) = l^{\ddagger}_{\psi}(\psi) + O_p(\sqrt{q}) + O_p\left(\frac{q}{m}\right)$$

and that $l_I(\psi)$ has score bias of order $O(\frac{q}{m})$. Note that this result holds for any choice of the density π. Also note that here ϕ_i can be a standard type of orthogonal parameter based on the expected information matrix, as discussed in Cox and Reid (1987) or the zero-score expectation parameter used in Severini (2007), and described in Section 2.2.

Related results hold for other likelihood-based quantities. For instance, let $\bar{\psi}$ denote the maximizer of $l_I(\psi)$ and let $\psi^{\ddagger}$ denote the maximizer of $l^{\ddagger}(\psi)$. Then, using the usual expansion for the maximum-likelihood-type estimators (e.g., Severini, 2000, Section 5.3), together with the fact that the marginal model is a standard one-parameter model,

$$\sqrt{mq}(\bar{\psi} - \psi) = \sqrt{mq}(\psi^{\ddagger} - \psi) + \bar{D}\sqrt{mq} + O_p\left(\frac{1}{mq}\right) + O_p\left(\frac{\sqrt{q}}{m^{\frac{3}{2}}}\right)$$

where

$$\bar{D} = \frac{1}{i^{\ddagger}} \sum_{i=1}^{q} D_i(\psi)$$

and $i^{\ddagger}$ denotes the expected information in the marginal model.

In general, $\bar{D} = O(1/m)$ and, hence,

$$\sqrt{mq}(\bar{\psi} - \psi) = \sqrt{mq}(\psi^{\ddagger} - \psi) + O_p\left(\frac{\sqrt{q}}{\sqrt{m}}\right) + O\left(\frac{1}{mq}\right) + O\left(\frac{\sqrt{q}}{m^{\frac{3}{2}}}\right).$$

It follows that $\bar{\psi}$ has the same asymptotic distribution as $\psi^{\ddagger}$ provided that $q/m = o(1)$. If $g(\lambda_i;\psi)/\pi(\lambda_i;\psi)$ is orthogonal to ψ, then

$$\bar{D} = O(m^{-2}) + O_p(1/\{\sqrt{m^3 q}\}).$$

It follows that $\bar{\psi}$ has the same asymptotic distribution as $\psi^{\ddagger}$ provided that $q/m^3 = o(1)$. Finally, if g is orthogonal to ψ, then

$$\bar{D} = O(m^{-2}) + O_p(1/(m\sqrt{q}));$$

it follows that $\bar{\psi}$ has the same asymptotic distribution as $\psi^{\ddagger}$ provided that $q/m^3 = o(1)$.

3.5 Signed square root integrated likelihood ratio statistics

We now consider likelihood-ratio-type statistics based on the integrated likelihood with g orthogonal to ψ. The asymptotic properties of such statistics under the marginal model can be obtained using the following approach. Since the marginal model is a standard one-parameter model, the asymptotic properties of signed likelihood ratio statistic in this model are well-understood. Using the relationship between the integrated likelihood and the likelihood based on the marginal model, we are able to determine the properties of the signed integrated likelihood ratio statistics under the marginal model.

Using the Laplace expansions in Section 3.3, it is straightforward to show that

$$\begin{aligned} l_I(\bar{\psi}) - l_I(\psi) =& \{l^{\ddagger}(\hat{\psi}^{\ddagger}) - l^{\ddagger}(\psi)\} + \{l^{\ddagger}(\bar{\psi}) - l^{\ddagger}(\hat{\psi}^{\ddagger})\} \\ &+ \{H(\bar{\psi}) - H(\psi)\} + \{B(\bar{\psi}) - B(\psi)\} \end{aligned}$$

where

$$H(\psi) = \sum_{i=1}^{q} H_i(\psi) = \sum_{i=1}^{q} \log \frac{g(\hat{\lambda}_{i\psi}; \psi)}{\pi(\hat{\lambda}_{i\psi}; \psi)}$$

and $B(\psi)$ is a remainder term that is order $O_p(q/m)$.

When g is orthogonal to ψ, we have seen that $\partial H_i(\psi)/\partial\psi$ has mean of order $O(1/m)$ so that $\partial H_i(\psi)/\partial\psi = O(\sqrt{q})$. Moreover, $\bar{\psi} = \hat{\psi}^{\ddagger} + O_p(1/\sqrt{m^2 q})$ provided that $q/m^3 = o(1)$. It follows that

$$l^{\ddagger}(\bar{\psi}) - l^{\ddagger}(\hat{\psi}^{\ddagger}) = O(m^{-1})$$

and

$$H(\bar{\psi}) - H(\psi) = O_p(m^{-\frac{1}{2}}).$$

Since $\bar{\psi} - \psi = O_p(1/\sqrt{mq})$,

$$B(\bar{\psi}) - B(\psi) = O_p\left(\sqrt{\frac{q}{m^3}}\right).$$

Hence,

$$l_I(\bar{\psi}) - l_I(\psi) = \{l^{\ddagger}(\hat{\psi}^{\ddagger}) - l^{\ddagger}(\psi)\} + O_p(m^{-\frac{1}{2}}) + O_p\left(\sqrt{\frac{q}{m^3}}\right). \tag{3.6}$$

Let $\bar{R}$ denote the signed likelihood ratio statistic based on l_I and let $R^{\ddagger}$ denote the signed likelihood ratio statistic based on $l^{\ddagger}$. Then

$$\bar{R} = R^{\ddagger} + O_p(m^{-\frac{1}{2}}) + O_p\left(\sqrt{\frac{q}{m^3}}\right)$$

provided that $q/m^3 = o(1)$. Since $R^{\ddagger}$ is asymptotically distributed according to a standard normal distribution under the marginal model, it follows that $\bar{R}$ is asymptotically standard normal under the marginal model provided that $q/m^3 = o(1)$. Note that this result holds for any choice of the density $\pi(\cdot;\psi)$.

Since the error in the normal approximation to the distribution of $R^\ddagger$ is of order $O_p(1/\sqrt{mq})$, the error in the normal approximation to the distribution of $\bar{R}$ has three components and is of order

$$O_p(1/\sqrt{mq}) + O_p(m^{-\frac{1}{2}}) + O_p\left(\sqrt{\frac{q}{m^3}}\right).$$

The first component is based on the overall sample size mq; this sample size drives the asymptotic normality of likelihood-based quantities such as the score statistic for ψ. The second component reflects the effect of the weight function used to construct L_I as it relates to the true values of the λ_i, as described by the density $\pi(\lambda_i;\psi)$. The third component reflects the error in the expansion of the integrated and marginal likelihood functions based on Laplace approximations; since the within-stratum error depends only on the within-stratum sample size, and the errors are compounded by summing across strata, this component depends on the relative magnitudes of q, m. In general, we expect the error in the normal approximation to the distribution of $\bar{R}$ to be small whenever m and m^3/q are relatively large.

In some cases, the error in the normal approximation to the distribution of $\bar{R}$ is of smaller order than that given above. For instance, suppose that $H_i(\psi)$ does not depend on ψ for each i; this occurs if $\pi(\lambda_i;\psi) = g(\lambda_i;\psi)$. It can also occur if there exists an orthogonal nuisance parameter ϕ such that $\hat{\phi}_{i\psi}$ does not depend on ψ and $g(\lambda_i;\psi)/\pi(\lambda_i;\psi)$ is a function of ϕ_i. Then

$$\bar{R} = R^\ddagger + O_p(m^{-1}) + O_p\left(\sqrt{\frac{q}{m^3}}\right)$$

provided that $q/m^3 = o(1)$.

These results can be compared to those for the usual signed likelihood ratio statistic R. Using results on the properties of the profile likelihood function and the maximum likelihood estimator in this setting (Sartori, 2003), it is straightforward to show that, if $q/m = o(1)$, then R is asymptotically normally distributed with error of order $O(m^{-\frac{1}{2}}) + O(q/m)$. If $q/m = o(1)$ does not hold, then R would not likely be asymptotically normal.

An important issue is the extent to which $\bar{R}$ depends on the weight function used in its construction. Let $g_1(\lambda_i;\psi)$ and $g_2(\lambda_i;\psi)$ denote two weight functions, each orthogonal to ψ and let $\bar{R}_1, \bar{R}_2$ denote the corresponding signed integrated likelihood ratio statistics. Then

$$\bar{R}_1 = \bar{R}_2 + O_p\left(\frac{1}{\sqrt{m}}\right) + O_p\left(\sqrt{\frac{q}{m^3}}\right).$$

Thus, provided that q is not too large relative to m, the difference between $\bar{R}_1$ and $\bar{R}_2$ is driven by the within-stratum sample size. For instance, if there is a large number of strata with relatively few observations with each stratum, then the choice of weight function may have an important effect on the the distributional accuracy of the signed integrated likelihood ratio statistic. On the other hand, if $q/m^3 = o(1)$, the choice of weight function is relatively unimportant.

3.6 Weight functions

Reconsidering expression (3.6). When q is large, even if $\pi(\lambda_i;\psi)$ and $g(\lambda_i;\psi)$ are orthogonal, or the parameter ϕ_i described in previous section is used, there is an effect of the choice of weight function, and it appears in

$$B(\bar{\psi}) - B(\psi) = \left\{Q(\bar{\psi}) - Q(\psi)\right\} - \left\{V(\bar{\psi}) - V(\psi)\right\} + O_p\left(\sqrt{\frac{q}{m^5}}\right),$$

through the term $\left\{Q(\bar{\psi}) - Q(\psi)\right\}$. Here $Q(\psi) = \sum_{i=1}^{q} Q_i(\psi)$ and $V(\psi) = \sum_{i=1}^{q} V_i(\psi)$.

If we study the order of term $B(\bar{\psi}) - B(\psi)$, it is led by quantity

$$(\bar{\psi} - \psi)Q_\psi(\psi) + (\bar{\psi} - \psi)V_\psi(\psi),$$

where, as usual, the subscript denotes derivative, in this case with respect to ψ.

The quantity $V_\psi(\psi)$ depends on the unknown distribution $\pi(\lambda;\psi)$, and so it can not be managed. On the contrary, we can act on $Q_\psi(\psi)$ and, in both summands, on the quantity $(\bar{\psi} - \psi)$. The bias of the estimator based on the integrated likelihood, indeed, depends on the choice of the weight function, always through the term $Q(\psi)$.

Theoretically, the best choice of the weight function is obviously $g(\lambda;\psi) = \pi(\lambda;\psi)$, since in this case $Q(\psi) = V(\psi)$ and the score bias for the integrated likelihood would be $O_p((mq)^{-1})$. Since $\pi(\lambda;\psi)$ is unknown, we have to find other solutions. The most conservative one is to force the term $Q(\psi)$ to be independent of ψ. In this way, we are sure not to increase the bias of the score to the score function beyond $V(\psi)$. Reconsidering expansion (3.1) and definition (3.2), it is easy to show that a weight function which achieves this result is the function g^* which solves in g the differential equation

$$1-\frac{5}{24}\frac{|l_{\lambda\lambda\lambda}|^2}{|l_{\lambda\lambda}|^3}+\frac{1}{8}\frac{|l_{\lambda\lambda\lambda\lambda}|}{|l_{\lambda\lambda}|^2}+\frac{1}{2}\frac{|l_{\lambda\lambda\lambda}|}{|l_{\lambda\lambda}|^2}\frac{g_\lambda}{g}-\frac{1}{2}\frac{1}{|l_{\lambda\lambda}|}\frac{g_{\lambda\lambda}}{g}=h, \tag{3.7}$$

being h an arbitrary function independent of ψ.

From a pratical point of view, it is sufficient to solve

$$-\frac{5}{24}\frac{|l_{\lambda\lambda\lambda}|^2}{|l_{\lambda\lambda}|^3}+\frac{1}{8}\frac{|l_{\lambda\lambda\lambda\lambda}|}{|l_{\lambda\lambda}|^2}+\frac{1}{2}\frac{|l_{\lambda\lambda\lambda}|}{|l_{\lambda\lambda}|^2}\frac{g_\lambda}{g}-\frac{1}{2}\frac{1}{|l_{\lambda\lambda}|}\frac{g_{\lambda\lambda}}{g}=0. \tag{3.8}$$

The main issue is that the function g^* can not depend on ψ, in order to maintain the orthogonality, and thus it is not always available. Note that using this weight function, the integrated log-likelihood agrees to order $O(q/m^2)$ with the approximate conditional log-likelihood, i.e. the approximate conditional likelihood can be seen as an integrated likelihood with a weight function which makes the terms of order higher than first in the Laplace approximation independent of ψ. Indeed, when (3.7) holds, the Laplace expansion of $l_I(\psi)$ is

$$l_I(\psi)=l_P(\psi)-\sum_{i=1}^{q}\log|-l_{\lambda\lambda}(\psi,\hat\lambda_\psi)|^{\frac{1}{2}}+O(q/m^2)=l_{AC}(\psi)+O(q/m^2). \tag{3.9}$$

A different solution, which should reduce the score bias, is to find a weight function g which solves only the part

$$\frac{1}{2}\frac{|l_{\lambda\lambda\lambda}|}{|l_{\lambda\lambda}|^2}\frac{g_\lambda}{g}-\frac{1}{2}\frac{1}{|l_{\lambda\lambda}|}\frac{g_{\lambda\lambda}}{g}=h \tag{3.10}$$

of equation (3.7). In this way, the term $Q(\psi)$ remains

$$1-\frac{5}{24}\frac{|l_{\lambda\lambda\lambda}|^2}{|l_{\lambda\lambda}|^3}+\frac{1}{8}\frac{|l_{\lambda\lambda\lambda\lambda}|}{|l_{\lambda\lambda}|^2},$$

which is the part in common with $V(\psi)$. The simplest way to solve (3.10) is to use a uniform weight function. It is worth noting that this choice, which seems the preferable one, provides us with a pretty easy way to proceed: after the orthogonalization step, it is sufficient to integrate the likelihood with respect to λ and to use it to construct the signed square root integrated likelihood ratio statistic.

3.6.1 Weight function for the original parameterization

The orthogonalization step is necessary to reduce the score bias from $O(q)$ to $O(q/m)$. From a practical point of view, this means that we have to reparameterize the model before doing the integration and obtain our integrated likelihood. A different choice can be to use a weight function which takes into account the non-orthogonal parameterization, following the idea of Cox and Reid (1993). The aim is to construct a weight function based on the original parameterization that would act like a uniform one in an orthogonal parameterization. The desired weight function is

$$g_P(\lambda;\psi) = \left|\frac{\partial\xi(\psi,\lambda)}{\partial\lambda}\right|^{-1}, \tag{3.11}$$

where $\xi(\psi,\lambda)$ is the solution of the partial differential equation

$$\frac{\partial\lambda(\xi,\psi)}{\partial\psi} = -i_{\lambda\lambda}^{-1} i_{\lambda\psi},$$

see Cox and Reid (1993, formula 2). It is worth noting that we do not need to find an explicit expression for $\lambda(\xi,\psi)$, because the weight function (3.11) requires only the computation of $\partial\xi/\partial\lambda$. Example 3.4 presents an instance of this construction.

3.7 Examples

Example 3.1 (Matched gamma pairs)
Let Y_{ij1} and Y_{ij2}, $i = 1,\ldots,q$, $j = 1,\ldots,m$, be independent exponential random variables with means ψ/λ_i and $\psi\lambda_i$, respectively, as in Example 4 in Sartori (2003). The parameter of interest is ψ, while there is a nuisance parameter, λ_i, for each stratum. Let $y_{i1} = \sum_j y_{ij1}$ and $y_{i2} = \sum_j y_{ij2}$, so

that $Y_{i1} \sim \text{Gamma}(m, \frac{\lambda_i}{\psi})$ and $Y_{i2} \sim \text{Gamma}(m, \frac{1}{\lambda_i \psi})$, the likelihood is

$$L(\psi, \lambda) = \psi^{-2mq} \exp\left\{ -\frac{1}{\psi} \sum_i \left(\lambda_i y_{i1} + \frac{y_{i2}}{\lambda_i} \right) \right\},$$

where $\lambda = (\lambda_1, \dots, \lambda_q)$. The constrained estimate of λ_i is $\hat{\lambda}_{i\psi} = \sqrt{y_{i2}/y_{i1}}$, which is independent of ψ, so $\hat{\lambda}_{i\psi} = \hat{\lambda}_i$ for every i. The estimate for ψ is $\hat{\psi} = \sum_i \sqrt{y_{i1} y_{i2}}/(mq)$ and the signed root likelihood ratio statistic is $R = \text{sgn}(\hat{\psi} - \psi)\sqrt{4mq \left[\log(\psi/\hat{\psi}) + (\hat{\psi} - \psi)/\psi \right]}$.

If we use a weight function $g(\lambda_i; \psi) = \lambda_i^{p-1}$, $p \in \mathbb{R}$, we obtain the integrated likelihood on the form

$$L_I(\psi) = \psi^{(-2mq+1/2)q} \prod_{i=1}^{q} K_p \left(\frac{2}{\psi \sqrt{y_{i1} y_{i2}}} \right),$$

where $K_p(\cdot)$ is the modified Bessel function of the third kind. This explains very well the effect of the choice of the weight function presented in Section 3.6. Besides the choice of a constant weight function, in this example we can compute $g^*(\lambda_i)$. Substituting the values in (3.8)

$$-\frac{\frac{5}{24}\left(\frac{6y_{i2}}{\psi\hat{\lambda}_{i\psi}^4}\right)^2}{\left(-\frac{2y_{i2}}{\psi\hat{\lambda}_{i\psi}^3}\right)^3} g + \frac{\frac{1}{8}\left(-\frac{24y_{i2}}{\psi\hat{\lambda}_{i\psi}^5}\right)}{\left(-\frac{2y_{i2}}{\psi\hat{\lambda}_{i\psi}^3}\right)^2} g + \frac{\frac{1}{2}\frac{6y_{i2}}{\psi\hat{\lambda}_{i\psi}^4}}{\left(-\frac{2y_{i2}}{\psi\hat{\lambda}_{i\psi}^3}\right)^2} g' - \frac{\frac{1}{2}}{\left(-\frac{2y_{i2}}{\psi\hat{\lambda}_{i\psi}^3}\right)} g'' = 0$$

and considering a generic $g = \lambda^\alpha$, we obtain

$$\alpha^2 + 2\alpha + \frac{3}{4} = 0,$$

that is solved by $\alpha = -1 \pm 1/2$. Hence g^* is of the form $\lambda_i^{-1\pm1/2}$, i.e. $p = \pm 1/2$.

Simulation study

We perform a simulation study replicating $B = 9000$ times the following procedure. Each time, we simulate, in $q = 1000$ strata, $m = 10$ pairs of observations y_{ij1} and y_{ij2} from exponential random variables with means ψ/λ_i and $\psi\lambda_i$, respectively. The nuisance parameters are generated from a χ^2 with 10 degrees of freedom random variable. The true value of ψ is 2.

Table 3.1: Matched gamma pairs: empirical coverage of signed root likelihood ratio statistics based on profile likelihood (R), integrated likelihoods with a generic weight function ($\bar{R}$), uniform weight function ($\bar{R}_0$) and weight function g^* ($\bar{R}_{g^*}$).

nominal	R	$\bar{R}$	$\bar{R}_{g^*}$	$\bar{R}_0$
0.01	0.885	0.001	0.010	0.009
0.025	0.946	0.004	0.024	0.024
0.05	0.973	0.009	0.047	0.046
0.1	0.987	0.022	0.093	0.093
0.5	0.999	0.237	0.478	0.478
0.9	1.000	0.712	0.897	0.897
0.95	1.000	0.826	0.951	0.951
0.975	1.000	0.895	0.976	0.975
0.99	1.000	0.951	0.991	0.991

Table 3.1 shows the empirical coverages of R, $\bar{R}$, $\bar{R}_{g^*}$, $\bar{R}_0$. Here we indicate with $\bar{R}_0$ the integrated likelihood with a constant weight function, while with $\bar{R}_{g^*}$ the one with the weight function which satisfies (3.7). The integrated likelihood $\bar{R}$ is computed with a weight function $g(\lambda) = \lambda^3$.

As we expected, in this example the usual statistic R does not perform well. Viceversa, the integrated version, $\bar{R}$, provides better coverages. They are not totally satisfactory, but neither so bad, taking into account that the weight function used, λ^3, is quite strange and extreme. Anyway, with a wise choice of the weight function, we achieve better results, as we can see by checking the empirical coverages of both $\bar{R}_{g^*}$ and $\bar{R}_0$.

Figure 3.1 shows graphically these behaviours. The distribution of R is quite far from the standard normal, while $\bar{R}$ is much closer. Finally, the distributions of $\bar{R}_{g^*}$ and $\bar{R}_0$ are indistinguishable and very close to the standard normal.

Example 3.2 (Gamma samples with common shape parameter)
Let Y_{ij}, $i = 1, \dots, q$, $j = 1, \dots, m$, be independent gamma random variables with shape parameter ψ and scale parameter $1/\lambda_i$, as in Barndorff-Nielsen (1996, Example 5.1). Writing $s = u - m\sum_{i=1}^{q} \log v_i$, where $u = \sum_{i=1}^{q}\sum_{j=1}^{m} \log y_{ij}$ and $v_i = \sum_{j=1}^{m} y_{ij}$ are the components of the sufficient

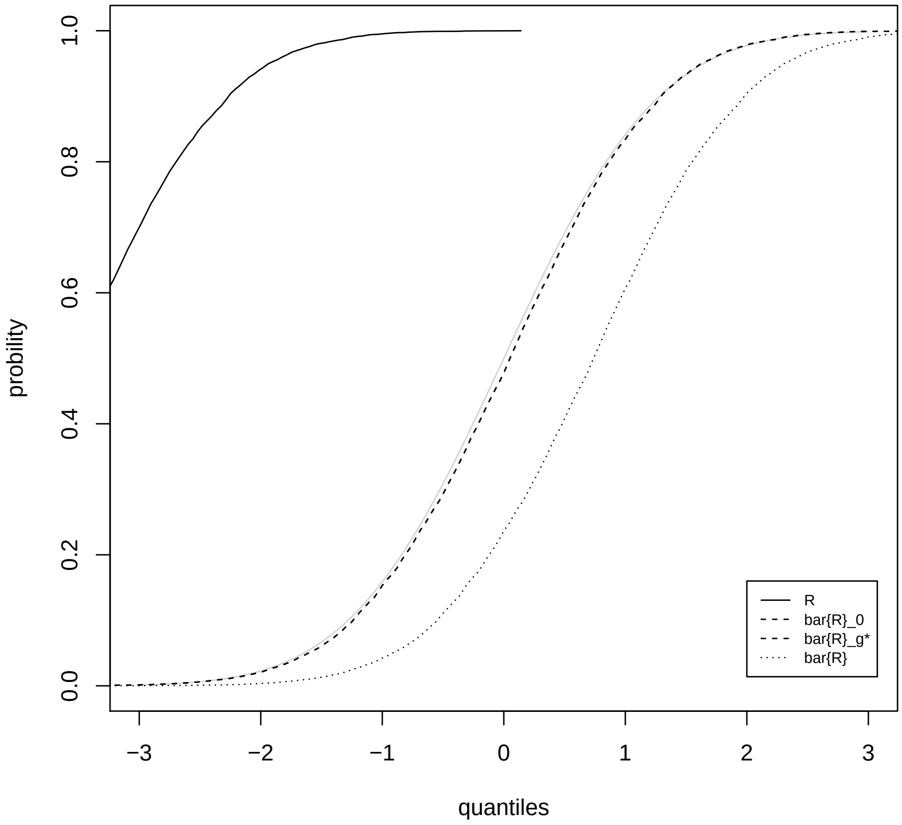

Figure 3.1: Matched gamma pairs: empirical distribution function of several signed root likelihood ratio statistics. Continuous grey line is the standard normal.

statistic, the conditional and profile log-likelihoods are

$$\begin{aligned} l_C(\psi) &= \psi s + q \log\Gamma(m\psi) - mq \log\Gamma(\psi), \qquad (3.12) \\ l_P(\psi) &= \psi s + mq\psi \log m\psi - mq\psi - mq \log\Gamma(\psi), \end{aligned}$$

while, if we use as a weight function the density function of an exponential random variable with mean 1, the integrated log-likelihood is

$$\begin{aligned} l_I(\psi) =& \psi \sum_{i=1}^{q} \sum_{j=1}^{m} \log y_{ij} - m\psi \sum_{i=1}^{q} \log(\sum_{j=1}^{m} y_{ij} + 1) + q \log\Gamma(m\psi + 1) - mq\psi \\ & - mq \log\Gamma(\psi). \qquad (3.13) \end{aligned}$$

In order to achieve orthogonality, we use the zero-score-expectation parameter, $\phi_i = \frac{\hat{\psi}\lambda_i}{\psi}$. This leads to the log-likelihood

$$l(\psi, \phi_1, \ldots, \phi_q) = \sum_{i=1}^{q} (\psi \sum_{j=1}^{m} \log y_{ij} + m\psi \log \frac{\phi_i \psi}{\hat{\psi}} - \frac{\phi_i \psi}{\hat{\psi}} \sum_{j=1}^{m} y_{ij} - m \log\Gamma(\psi)).$$

The integrated log-likelihood derives from it by integration, using a weight function independent of ψ. For the generic comparison $\bar{R}$ we use the weight function $g(\phi) = \phi^2$, while, as in previous example, with $\bar{R}_0$ we describe the signed square root likelihood ratio statistics with integrated likelihood using a constant weight function and with $\bar{R}_{g^*}$ the one based on a integrated likelihood with weight function $g^*(\phi_i) = \phi_i^{(-1 \pm \sqrt{1/3})/2}$. The latter is the solution of (3.8),

$$\frac{5}{24} \frac{4m^2\psi^2}{\phi_i^6} \frac{\phi_i^6}{m^3\psi^3} g - \frac{1}{8} \frac{6m\psi}{\phi_i^4} \frac{\phi_i^4}{m^2\psi^2} g + \frac{1}{2} \frac{2m\psi}{\phi_i^3} \frac{\phi_i^4}{m^2\psi^2} g' + \frac{1}{2} \frac{\phi_i^2}{m\psi} g'' = 0.$$

In this example, the conditional likelihood and the integrated likelihood with uniform weight function are equivalent. In fact, denoting the latter by $\bar{L}_0$, we have

$$\begin{aligned} \bar{L}_0(\psi) =& \prod_{i=1}^{q} \int_{\mathbb{R}^+} e^{\psi \sum_{j=1}^{m} \log y_{ij} + m\psi \log \frac{\phi_i \psi}{\hat{\psi}} - \frac{\phi_i \psi}{\hat{\psi}} \sum_{j=1}^{q} y_{ij} - m \log\Gamma(\psi)} d\phi_i \\ =& \prod_{i=1}^{q} \left[e^{\psi \sum_{j=1}^{m} \log y_{ij} - m \log\Gamma(\psi)} \left(\frac{\psi}{\hat{\psi}}\right)^{m\psi} \int_{\mathbb{R}^+} \phi_i^{m\psi} e^{-\frac{\phi_i \psi}{\hat{\psi}} \sum_{j=1}^{q} y_{ij}} d\phi_i \right]. \end{aligned}$$

We recognize the kernel of a Gamma$(m\psi + 1, \frac{\psi}{\hat{\psi}} \sum_{j=1}^{q} y_{ij})$, and

$$\bar{L}_0(\psi) = \prod_{i=1}^{q} \left[e^{\psi \sum_{j=1}^{m} \log y_{ij} - m \log \Gamma(\psi)} \frac{\hat{\psi}}{\psi} \left(\frac{1}{\sum_{j=1}^{q} y_{ij}} \right)^{m\psi+1} \Gamma(m\psi + 1) \right]$$

$$\prod_{i=1}^{q} e^{\psi(\sum_{j=1}^{m} \log y_{ij} - m \log \sum_{j=1}^{q} y_{ij}) - m \log \Gamma(\psi) + \log \Gamma(m\psi+1) - \log \psi}.$$

Since $\Gamma(m\psi + 1) = m\psi\Gamma(m\psi)$, we obtain

$$\bar{L}_0(\psi) = \prod_{i=1}^{q} e^{\psi(\sum_{j=1}^{m} \log y_{ij} - m \log \sum_{j=1}^{q} y_{ij}) - m \log \Gamma(\psi) + \log \Gamma(m\psi)},$$

and its logarithm is exactly (3.12). It is worth noting that, since the conditional and marginal likelihood for inference on the shape parameter of a Gamma distribution are equivalent, 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 Salvan, 1997, Example 7.29). As a final remark, we note that the modified profile likelihood in this example is not equivalent to the conditional likelihood (Sartori, 2003, Example 2).

Simulation study

Also for this example, we perform a simulation study replicating $B = 8000$ times the following procedure. Each time, we simulate, in $q = 1000$ strata, $m = 10$ observations y_{ij} from gamma random variables with shape parameter ψ and scale parameter $1/\lambda_i$. The nuisance parameters are generated from a χ^2 with 10 degrees of freedom. The true value of ψ is 2.

Table 3.2 shows the empirical coverages of several signed root likelihood ratio statistics. The empirical coverages of R and R_I are, as expected, very poor. Here, with R_I we denote the signed square root integrated likelihood ratio statistic based on (3.13). Some non-negligible discrepancies are present between the nominal and the empirical coverages for $\bar{R}$, while, choosing wisely the weight function, we can see how these discrepancies disappear, leading to empirical coverages very close to the nominal ones, both using an uniform weight function, which is equivalent to the conditional likelihood, or g^*. Figure 3.2 shows the empirical distributions. The distribution of R_C and $\bar{R}_0$, in continuous line, is indistinguishable from the standard

Table 3.2: Gamma with common shape parameter: empirical coverage of signed root likelihood ratio statistics, based on profile likelihood (R), conditional likelihood (R_C), integrated likelihood with conjugate weight function (R_I), integrated likelihoods with orthogonalization step and generic weight function ($\bar{R}$), uniform weight function ($\bar{R}_0 = R_C$) and weight function g^* ($\bar{R}_{g^*}$).

nominal	R	$R_C = \bar{R}_0$	R_I	$\bar{R}$	$\bar{R}_{g^*}$
0.01	0.000	0.009	1.000	0.045	0.008
0.025	0.000	0.024	1.000	0.095	0.020
0.05	0.000	0.047	1.000	0.156	0.043
0.1	0.000	0.098	1.000	0.258	0.085
0.25	0.000	0.240	1.000	0.489	0.222
0.5	0.000	0.498	1.000	0.740	0.471
0.75	0.000	0.745	1.000	0.906	0.727
0.9	0.000	0.896	1.000	0.973	0.886
0.95	0.000	0.949	1.000	0.988	0.943
0.975	0.000	0.975	1.000	0.996	0.971
0.99	0.000	0.990	1.000	0.999	0.988

normal. Also the statistics $\bar{R}_{g^*}$, in dotted line, is very close to standard normal distribution. On the contrary, both R and R_I are very far from the standard normal, especially the latter, whose distribution is outside the plotting region. Recalling the fact that also this setting is pretty extreme, it is worth noting that the behaviour of the integrated signed root likelihood ratio statistic with the zero-score expectation parameterization even using the unusual weight function λ^2 is better than the usual R.

We perform a simulation study also to understand the effect of the magnitude of m and q on the normal approximation of the proposed statistics. The results are provided in Table 3.3. As expected, all the empirical coverage are closer to the nominal one increasing m and decreasing q. However, it is worth noting that the statistics based on integrated likelihood with uniform or g^* weight function perform very well also in case of small m and large q. This may be a particularity of this example, in which there is exact (uniform) or very close (g^*) agreement between integrated and conditional likelihoods.

Example 3.3 (Exponential Regression)
Consider the exponential regression model with one covariate, where $E[Y_i] = \lambda_i \exp\{-\psi z_i\}$, with $\sum_j z_{ij} = 0$, as in Cox and Reid (1987, Example 4.2.2), but with a nuisance parameter for each stratum. The condition $\sum_j z_{ij} = 0$

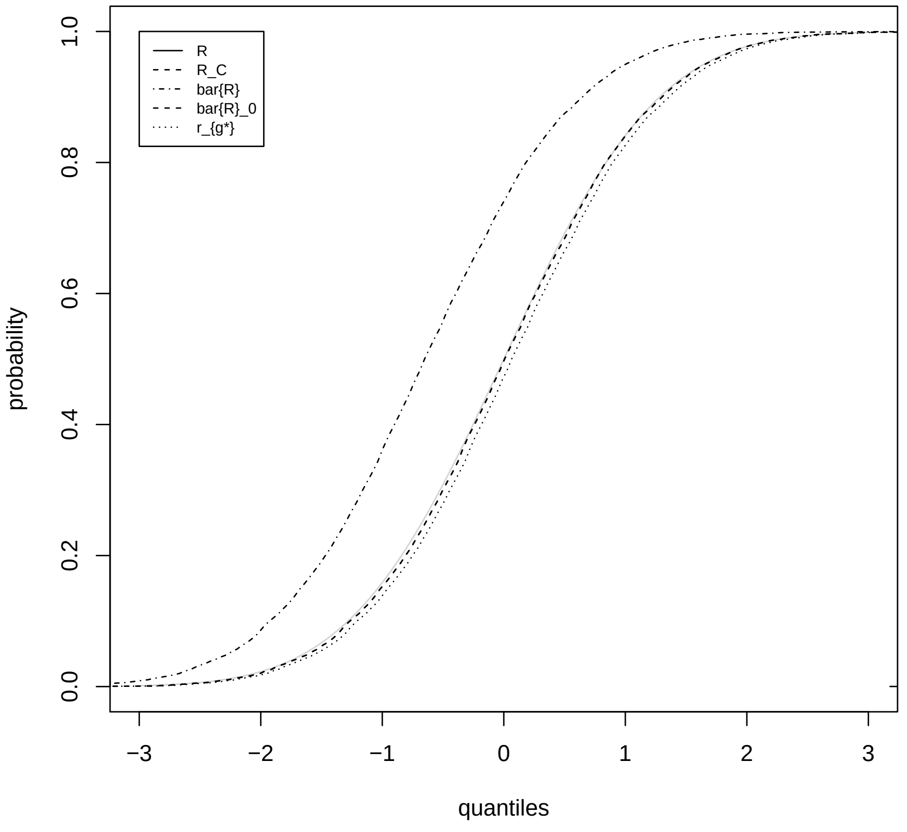

Figure 3.2: Gamma with common shape parameter: empirical distribution function of several signed root likelihood ratio statistics.

Table 3.3: Gamma with common shape parameter: empirical coverage probabilities of the 0.95 quantile of signed square root likelihood ratio statistics based on profile likelihood (R), conditional likelihood (R_C), integrated likelihood with conjugate weight function (R_I), integrated likelihoods with orthogonalization step and generic weight function, namely $g(\phi) = \phi^3$ ($\bar{R}$), uniform weight function ($\bar{R}_0 = R_C$) and weight function g^* ($\bar{R}_{g^*}$). Simulations are performed with 8000 replications each and with various values of m and q.

q	m	R	$R_C = \bar{R}_0$	R_I	$\bar{R}$	$\bar{R}_{g^*}$
100	5	0.069	0.949	1.000	0.999	0.944
	10	0.295	0.952	1.000	0.985	0.950
	20	0.544	0.948	1.000	0.966	0.947
200	5	0.003	0.949	1.000	1.000	0.943
	10	0.083	0.950	1.000	0.993	0.950
	20	0.325	0.948	1.000	0.972	0.947
1000	5	0.000	0.953	1.000	1.000	0.933
	10	0.000	0.949	1.000	1.000	0.943
	20	0.001	0.951	1.000	0.989	0.950

makes ψ and λ orthogonal. Here $i = 1, \ldots, q$, $j = 1, \ldots, m$. The log-likelihood is

$$l(\psi, \lambda) = \sum_{i=1}^{q} \left(-m \log \lambda_i - \lambda_i^{-1} \sum_{j=1}^{m} y_{ij} \exp\{\psi z_{ij}\} \right) .$$

It is straightforward to find $\hat{\lambda}_{i\psi} = m^{-1} \sum_j y_{ij} \exp\{\psi z_{ij}\}$ and that $\hat{\psi}$ satisfies $\sum_{i,j} y_{ij} z_{ij} \exp\{\psi z_{ij}\} = 0$.

The signed square root profile likelihood ratio statistic is

$$\begin{aligned} R = & \operatorname{sgn}(\hat{\psi} - \psi) \sqrt{2(-m \sum_{i=1}^{q} \log \sum_{j=1}^{m} y_{ij} \exp\{\hat{\psi} z_{ij}\} + m \sum_{i=1}^{q} \log \sum_{j=1}^{m} y_{ij} \exp\{\psi z_{ij}\})} \\ = & \operatorname{sgn}(\hat{\psi} - \psi) \sqrt{2 \left[m \sum_{i=1}^{q} \log \left(\hat{\lambda}_{i\psi} / \hat{\lambda}_{i\hat{\psi}} \right) \right]} \end{aligned}$$

The general form of the integrated likelihood is

$$L_I(\psi) = \prod_{i=1}^{q} \int \lambda_i^{-m} \exp\left\{ -\frac{\sum_j y_{ij} \exp\{\psi z_{ij}\}}{\lambda_i} \right\} g(\lambda_i; \psi) d\lambda_i.$$

Among the possible weight functions, we can choose the $g(\lambda_i; \psi)$ which

Table 3.4: Exponential regression: empirical coverage of signed root likelihood ratio statistics, based on profile likelihood (R), integrated likelihood with standard normal weight function ($\bar{R}$), with uniform weignt function ($\bar{R}_0$) and with weight function g^* ($\bar{R}_{g^*}$).

nominal	R	$\bar{R}$	$\bar{R}_0$	$\bar{R}_{g^*}$
0.01	0.014	0.031	0.007	0.015
0.025	0.033	0.054	0.027	0.032
0.05	0.061	0.089	0.047	0.064
0.1	0.110	0.137	0.099	0.111
0.25	0.245	0.273	0.234	0.249
0.5	0.511	0.503	0.511	0.511
0.75	0.760	0.734	0.767	0.757
0.9	0.910	0.879	0.921	0.905
0.95	0.949	0.923	0.957	0.946
0.975	0.971	0.957	0.981	0.971
0.99	0.989	0.972	0.993	0.988

solves (3.8),

$$\frac{10}{3m}g - \frac{9}{4m}g + \frac{2\lambda}{m}g' + \frac{\lambda^2}{2m}g'' = 0,$$

i.e. $g^* = \lambda^{\frac{-3\pm\sqrt{9-8(13/12)}}{2}}$.

Simulation study

We perform a simulation study replicating $B = 9000$ times the following procedure. Each time, we simulate in $q = 1000$ strata, $m = 10$ observations y_{ij}, while, in each stratum, z_{ij} are simulated from a standard normal and then recentred in order to satisfy the constraint $\sum_j z_{ij} = 0$. The nuisance parameters are generated from an exponential random variable with mean 1. The true value of ψ is 2.

As we can see in Table 3.4, the empirical coverages of integrated signed root likelihood ratio statistic are very close to the nominal ones for each choice of weight function. Here, we use a standard normal weight function for $\bar{R}$, while, as in the previous examples, $\bar{R}_0$ and $\bar{R}_{g^*}$ denote the signed square root likelihood ratio statistics based on an integrated likelihoods with uniform and g^* weight function, respectively. In this example also the statistic based on profile likelihood performs well: this is due to the fact that here the profile likelihood approximates the modified profile likelihood (see Severini, 2000, Example 9.7).

Example 3.4 (Statistical detection of a noisy signal)
Let us consider the problem of detecting a signal in the presence of background noise. We focus on the example handled by Davison and Sartori (2008), while to deepen the statistical issues involved in this kind of problems, we can refer to Mandelkern (2002) and to Fraser et al. (2004). The highly idealized version that we consider is a q strata model where, in each stratum, the observation is a realization of $Y_i = (Y_{1i}, Y_{2i}, Y_{3i})$, $i = 1, \dots, q$, where the three components are independent Poisson variables with means $(\gamma_i\psi + \beta_i, \beta_i t_i, \gamma_i u_i)$ respectively. Here Y_{1i} represents the main measurement, Y_{2i} and Y_{3i} are subsidiary background and efficiency measurements respectively, while t_i and u_i are known constants.

Reparameterizing the model in terms of ψ, γ_i and $\lambda_i = \beta_i/\gamma_i$, and conditioning on $S_i = Y_{1i}+Y_{2i}+Y_{3i}$, we obtain a trinomial density for (Y_{1i}, Y_{2i}, Y_{3i}) independent of γ_i. Up to multiplicative constants, the corresponding likelihood is

$$L(\psi, \lambda_i) = \prod_{i=1}^{q} \frac{(\psi + \lambda_i)^{y_{1i}} \lambda_i^{y_{2i}}}{(\psi + \lambda_i + u + \lambda_i t)^{s_i}}$$

An orthogonal parameter is a solution of the equation

$$\frac{\partial \lambda_i(\psi, \xi_i)}{\partial \psi} = \frac{\lambda_i(\psi, \xi_i)(\psi t_i - u_i)}{\psi t(\psi + u_i) + \lambda_i(\psi, \xi_i) u_i (1 + t_i)},$$

for instance

$$\xi_i(\psi, \lambda_i) = t_i \log \lambda_i + \log(\lambda_i + \psi) - (1 + t_i) \log(\psi + \lambda + u_i + \lambda t_i),$$

see Davison and Sartori (2008). It is impossible to express λ_i explicitly as a function of ψ and ξ_i, but in order to use formula (3.11) we only need to compute

$$\frac{\partial \xi_i}{\partial \lambda_i} = \frac{t_i u_i \lambda_i + t_i \psi^2 + t_i u_i \psi + u_i \lambda_i}{\lambda_i(\lambda_i \psi)(\psi + \lambda_i + u_i + t_i \lambda_i)}.$$

This gives

$$L_I(\psi) = \prod_{i=1}^{q} \int_0^{\infty} \frac{\psi(\psi + \lambda_i)^{y_{1i}} \lambda_i^{y_{2i}+2}}{(\psi + \lambda_i + u + \lambda_i t)^{s_i - 1}} \frac{1}{t_i u_i \lambda_i + t_i \psi^2 + t_i u_i \psi + u_i \lambda_i} d\,\lambda_i.$$

Table 3.5: Statistical detection of a noisy signal: empirical coverage of signed root likelihood ratio statistics based on profile likelihood (R), integrated likelihood with a generic weight function ($\bar{R}$), integrated likelihood with zero-score expectation reparameterization and uniform weight function ($\bar{R}_0$) and integrated likelihood in orginal parameterization and weight function 3.11.

nominal	R	$\bar{R}$	$\bar{R}_0$	R_{wf}
0.01	0.056	0.012	0.011	0.010
0.025	0.074	0.029	0.026	0.026
0.05	0.100	0.057	0.054	0.054
0.1	0.150	0.113	0.106	0.108
0.25	0.297	0.274	0.260	0.276
0.5	0.531	0.523	0.509	0.538
0.75	0.760	0.767	0.756	0.778
0.9	0.896	0.909	0.902	0.917
0.95	0.943	0.953	0.950	0.960
0.975	0.967	0.978	0.975	0.980
0.99	0.982	0.992	0.990	0.992

Simulation study

We simulated $B = 9000$ times observations from a $q = 10$ strata process, setting the parameters and the constants as in Davison and Sartori (2008, Table 3): $\psi = 2$, $\beta = (0.20, 0.30, 0.40, \ldots, 1.10)$, $\gamma = (0.20, 0.25, 0.30, \ldots, 0.65)$, $t = (15, 17, 19, \ldots, 33)$ and $u = (50, 55, 60, \ldots, 95)$. In Table 3.5 we report the empirical coverage probabilities of signed root likelihood ratio statistics based on profile likelihood (R), integrated likelihoods with uniform weight function without the orthogonalization step ($\bar{R}$) and with zero-score expectation reparameterization ($\bar{R}_0$), and, finally, the integrated likelihood with the weight function (3.11).

The empirical coverages of $\bar{R}_0$ and of R_{wf} are very close to each other and to the nominal value. $\bar{R}$ performs pretty well. Anyway, also in this example in which $\bar{R}$ behaves relatively well, its coverages are worse than the ones of the signed ratio statistics based on orthogonality. In this example, R_{wf}, based on the expected information matrix reparameterization through weight function (3.11) and $\bar{R}_0$, based on zero-score expectation parameterization, have a similar behaviour, and outperform the one based on profile likelihood, especially in the tails. The empirical distribution of the various statistics is reported in Figure 3.3.

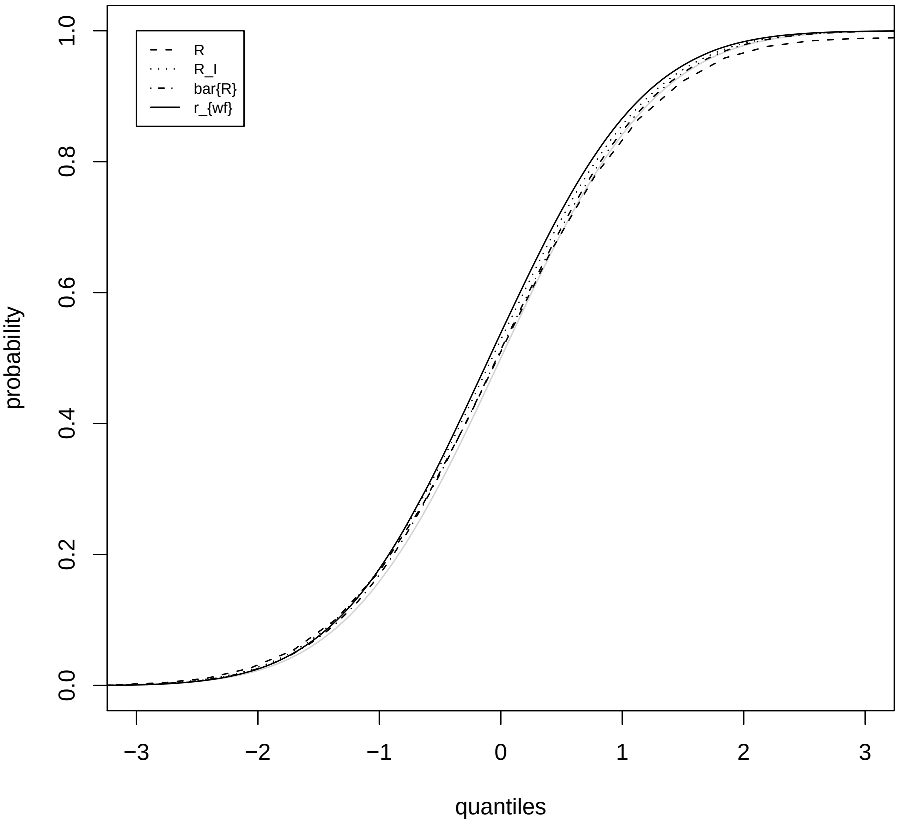

Figure 3.3: Statistical detection of a noisy signal: empirical distribution function of several signed root likelihood ratio statistics.

Example 3.5 (Gaussian with common variance example)
We consider an application of the canonical Neyman-Scott example (Neyman and Scott, 1948). Consider y_{ij}, $i = 1, \dots, q$, $j = 1, \dots, m$, realizations of q Gaussian random variables Y_i with common variance σ^2 and stratum-dependent mean λ_i. The variance σ^2 is the parameter of interest and $\lambda = (\lambda_1, \dots, \lambda_q)$ is the nuisance parameter. We consider below a special case where $\lambda_i = \lambda_i(\beta_i; x)$. Here β_i is three-dimensional and x represents a covariate. We consider the dataset presented in Appendix A.13 of Pinheiro and Bates (2000) concerning the growth of Loblolly trees. It contains information about $q = 14$ trees with different seed sources. For every tree, the dataset reports the height with respect to $m = 6$ different ages, namely $x = \{3, 5, 10, 15, 20, 25\}$. We assume that $Y_{ij} = \lambda_i(\beta_i; x_j) + \sigma\epsilon_{ij}$, where ϵ_{ij} are independent standard random variables and λ_i is

$$\lambda_i(\beta_i; x) = \beta_{1i} + (\beta_{2i} - \beta_{1i}) \exp\{-e^{\beta_{3i}} x\}.$$

Besides the usual profile maximum likelihood estimator ($\hat{\sigma}^2$) and the modified profile maximum likelihood estimator ($\hat{\sigma}^2_{MP}$, see Sartori (2003)), we compute the maximum integrated likelihood estimators using as weight functions both a multivariate standard normal ($\bar{\sigma}^2$) and a multivariate uniform distribution ($\bar{\sigma}^2_0$). The values obtained are:

Estimator	$\hat{\sigma}^2$	$\hat{\sigma}^2_{MP}$	$\bar{\sigma}^2$	$\bar{\sigma}^2_0$
Value	0.245	0.490	0.489	0.490
0.95 CI	(0.184;0.337)	(0.329;0.777)	(0.345;0.724)	(0.330,0.773)

The integrated likelihood based on uniform weight function gives an estimate very close to $\hat{\sigma}^2_{MP}$. Also the use of a Gaussian weight function in normal regression model is acceptable, leading to a point estimate very close to $\bar{\sigma}^2_0$. Figure 3.4 shows the relative log-likelihoods together with 0.95 confidence intervals. The two integrated log-likelihoods and the modified profile log-likelihood are quite close to each other. In particular, the integrated log-likelihood based on uniform weight function and the modified profile log-likelihood are almost indistinguishable in the plot and lead to pratically identically confidence interval. In particular, a simulation study performed by Sartori (2003) gave a coverage probability for the confidence interval based on the modified profile log-likelihood very close to the nominal one (0.951 versus 0.95, with the true value $\sigma^2 = 0.5$).

3.8 Discussion

In this chapter we compared the signed square root likelihood ratio statistic based on an integrated likelihood with the standard R based on the profile likelihood.

The simulation results show that $\bar{R}$, after an orthogonalization step and using a suitable weight function, has a good behaviour even in very extreme settings with a large number of strata and few observations in each stratum. In particular, it is worth noting that the best performances have been obtained with a relatively easy choice for the weight function, namely a constant. In Section 3.6 we provided a first hint to explain why this kind of weight function seems to work so well, but further investigation is recommended. It is worth noting that, in Example 3.2, the signed square root likelihood ratio statistics based on integrated likelihood with constant weight function is equivalent to the one based on the conditional likelihood. This seems to support our choice. Anyway, we have seen that, even using a generic weight function (for instance, the pretty uncommon $g = \lambda^3$ and $g = \lambda^2$ in Example 3.1 and 3.2 respectively), $\bar{R}$ outperforms the standard R.

In literature other statistics has been studied, based on higher order asymptotics, such as the directed modified profile likelihood ratio statistic R_{MP} and the modified directed profile likelihood ratio statistic R^* (Barndorff-Nielsen, 1986). The similarities between $\bar{R}$ and R^* was studied in the standard asymptotic setting by Severini (2010). In particular, in presence of orthogonal parameters, with a suitable choice of weight function, they agree to second order.

In the two-index asymptotics framework, the close relation between these two statistics still holds, and in the orthogonal parameter case they agree to order $O(\max\{1/\sqrt{mq}; 1/\sqrt{m}; \sqrt{q/m^3}\})$, provided that the weight function is independent on ψ. To this end, let us consider the relation between $\bar{R}$ and R_{MP}, since in the two-index asymptotics R^* and R_{MP} differ from each other by a quantity of order $O_p(1/\sqrt{mq})$ (Sartori et al., 1999). We have

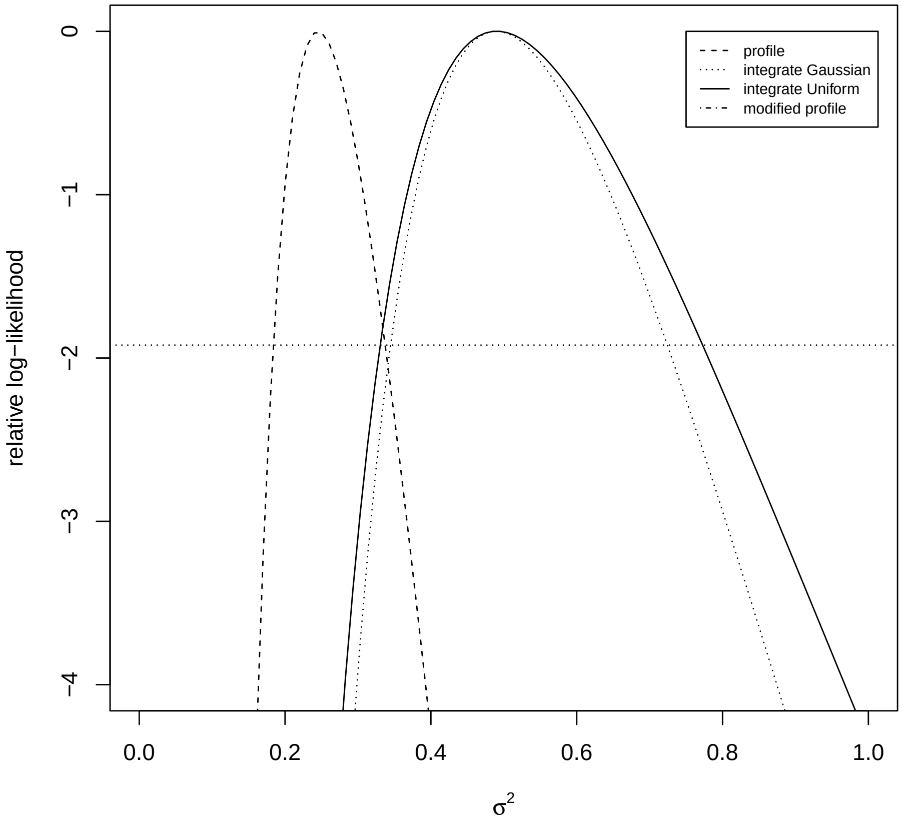

Figure 3.4: Loblolly data: relative log-likelihoods and confidence intervals for σ, based on profile log-likelihood (dashed line), modified profile log-likelihood (dot-dashed line, undistinguishable from continuous line), integrated log-likelihoods with Gaussian (dotted line) and uniform (continuous line) weight function.

already seen in Section 3.5 that

$$\bar{R} = R^{\ddagger} + O_p(1/\sqrt{m}) + O_p(\sqrt{q/m^3}),$$

while, for the directed modified profile ratio statistic, we can follow a similar argument. As a first step, the Laplace expansion gives

$$\begin{aligned} l_{MP}(\hat{\psi}_{MP}) - l_{MP}(\psi) =& \{l^{\ddagger}(\hat{\psi}^{\ddagger}) - l^{\ddagger}(\psi)\} + \{l^{\ddagger}(\hat{\psi}_{MP}) - l^{\ddagger}(\hat{\psi}^{\ddagger})\} + \\ & - \sum_{i=1}^{q} \log \frac{\pi(\hat{\lambda}_i^{\ddagger}; \hat{\psi}^{\ddagger})}{\pi(\hat{\lambda}_{i\psi}; \psi)} + O_p(q^{1/2} m^{-3/2}). \end{aligned}$$

The order of the summand $\sum_{i=1}^{q} \log(\pi(\hat{\lambda}_i^{\ddagger}; \hat{\psi}^{\ddagger})/\pi(\hat{\lambda}_{i\psi}; \psi))$ is $O_p(1/\sqrt{m})$, while the order of $\{l^{\ddagger}(\hat{\psi}_{MP}) - l^{\ddagger}(\hat{\psi}^{\ddagger})\}$ is $O_p(1/m)$. In the orthogonal parameter case, indeed, since $l_{AC}(\psi)$ approximates $l_{MP}(\psi)$ to second order, we can apply the argument seen for $\bar{\psi}$ in the case of a weight function orthogonal to ψ and we can state that $(\hat{\psi}_{MP} - \hat{\psi}^{\ddagger}) = O(1/\sqrt{m^2 q})$. As a consequence,

$$R_{MP} = R^{\ddagger} + O_p(1/\sqrt{m}) + O_p(\sqrt{q/m^3}),$$

and

$$\bar{R} = R^{*} + O_p(1/\sqrt{mq}) + O_p(1/\sqrt{m}) + O_p(\sqrt{q/m^3}).$$

The above relations provide an alternative proof to the asymptotic standard normal distribution of $\bar{R}$ and shows that, in the two-index asymptotic setting, $\bar{R}$ can be used to approximate R^{*}, especially when there is enough information in each stratum. As R_{MP}, the signed square root integrated likelihood statistic is able to recover the loss of information due to the presence of nuisance parameters in a way very similar to R^{*}, but it fails to consider the issues due to the lack of information. In order to do this, we should introduce some modification, as in Severini (2010). Finally, it is worth noting that the condition which assures the agreement between $\bar{R}$ and R^{*} is $q/m^3 = o(1)$, exactly the same which assures the convergence of the distribution of $\bar{R}$ to the standard normal distribution.

Chapter 4

Equivalence between integrated and conditional likelihood

4.1 Introduction

In this chapter we study the similarities and the possible equivalence between conditional and integrated likelihood. As stated in Rice (2008), their behaviours are really close to each other in many situations, and some common ground between the two inferential tools seems to exist. In Section 4.2 we take into consideration integrated likelihoods with a weight function that depends only on the parameter of interest, while in Section 4.3 we extend the range of possible weight functions allowing them to depend on an hyperparameter.

The idea presented in Section 4.2 is simple. Focusing on the exponential family framework, we provide a weight function which assures the equivalence between conditional and integrated likelihood. In Section 4.3 we consider cases where there is a discrete sufficient statistic for the nuisance parameter with finite support, and we investigate the relation between conditional and integrated approach focusing our attention on the hyperparameter, such as in Rice (2004, 2008).

4.2 Weight function that depends on the parameter of interest

Let Y be a random variable with density $p(y;\psi,\lambda)$ and consider a nonnegative weight function $g(\lambda;\psi)$, not necessarily a density function. In the presence of a sufficient statistic S for the nuisance parameter λ, the integrated likelihood (2.4) can be written as

$$L_I(\psi) = \int_\Lambda p(y;\psi,\lambda)g(\lambda;\psi)\,d\lambda = p_{Y|S=s}(y;\psi,s)\int_\Lambda p_S(s;\psi,\lambda)g(\lambda;\psi)\,d\lambda, \tag{4.1}$$

where Λ, independent of ψ, is the parameter space for λ. Note that the conditional density is taken out of the integral since it does not depend on λ.

The main idea in Rice (2004, 2008) is to force the factor

$$\int_\Lambda p_S(s;\psi,\lambda)g(\lambda;\psi)\,d\lambda \tag{4.2}$$

to be independent of ψ, in order to have all the information about the parameter of interest in the conditional density, so that $L_I(\psi)$ is equivalent to the conditional likelihood. This result is achieved through the choice of a suitable weight function.

Let us consider an exponential family framework, i.e. let the density function be of the form

$$p(y;\psi,\lambda) = \exp\{t\psi + s\lambda - K(\psi,\lambda)\}h(t,s),$$

where $K(\psi,\lambda)$ is the cumulant generating function of Y and t, s are function of y (see, for instance, Pace and Salvan, 1997, Section 5). Denoting with $p_0(s)$ the marginal density of S when $(\psi,\lambda) = (0,0)$ and with $M_s(\psi)$ the conditional moment generating function of t given $S = s$,

$$p_S(s;\psi,\lambda) = \exp\{\lambda s - K(\psi,\lambda)\}M_s(\psi)p_0(s). \tag{4.3}$$

Therefore, (4.2) can be written as

$$\int_\Lambda \exp\{\lambda s - K(\psi,\lambda)\}M_s(\psi)p_0(s)g(\lambda;\psi)\,d\lambda.$$

We can see that the only quantities dependent on ψ are $e^{-K(\psi,\lambda)}$ and $M_s(\psi)$. In order to have (4.2) independent of ψ, it is therefore sufficient to set

$$g(\lambda;\psi) = \frac{e^{K(\psi,\lambda)}}{M_s(\psi)}p_\lambda(\lambda) = \frac{M(\psi,\lambda)}{M_s(\psi)}p_\lambda(\lambda), \tag{4.4}$$

where $M(\psi,\lambda)$ is the moment generating function of Y and $p_\lambda(\lambda)$ is a generic function independent of ψ.

Using (4.4), indeed, the integrated likelihood becomes

$$L_I(\psi) = p_{Y|S=s}(y;\psi,s)\int_\Lambda e^{\lambda s}p_\lambda(\lambda)d\lambda \propto L_C(\psi).$$

Example 4.1 (Log-odds ratio)
Let Y_1 and Y_2 be two independent Bernoulli random variables with success probabilities p_1 and p_2. Here we consider $\psi = \log\{p_2(1-p_1)/[p_1(1-p_2)]\}$ as the parameter of interest and $\lambda = \log\{p_1/(1-p_1)\}$ as the nuisance parameter. A sufficient statistic has components $s = y_1 + y_2$ and $t = y_2$. The density

can be written as

$$p_{T,S}(t,s;\psi,\lambda) \propto \exp\{t\psi + s\lambda - [\log(1+e^{\lambda}) + \log(1+e^{\psi+\lambda})]\}.$$

Here

$$M(\psi,\lambda) = (1+e^{\lambda})(1+e^{\psi+\lambda}),$$

while with some algebra (see Pace and Salvan, 1997, example 5.12) we obtain

$$M_s(\psi) = \sum_{t'=max(0,s-1)}^{min(1,s)} e^{\psi t'}.$$

With these quantities, it is straightforward to write the weight function (4.4) as,

$$g(\lambda;\psi) = \frac{(1+e^{\lambda})(1+e^{\psi+\lambda})}{\sum_{t'=max(0,s-1)}^{min(1,s)} e^{\psi t'}} p_{\lambda}(\lambda).$$

The integrated likelihood is therefore

$$\begin{aligned} L_I(\psi) &= \int_{\mathbb{R}} \frac{e^{\lambda s+\psi y_2}}{(1+e^{\lambda})(1+e^{\lambda+\psi})} g(\lambda;\psi) d\lambda \\ &= \int_{\mathbb{R}} \frac{e^{\lambda s+\psi y_2}}{(1+e^{\lambda})(1+e^{\lambda+\psi})} \frac{(1+e^{\lambda})(1+e^{\lambda+\psi})}{\sum_{t'=max(0,s-1)}^{min(1,s)} e^{\psi t'}} p_{\lambda}(\lambda) d\lambda \\ &\propto \frac{e^{\psi y_2}}{\sum_{t'=max(0,s-1)}^{min(1,s)} e^{\psi t'}}. \end{aligned}$$

Since this expression is constant both for $S=0$ and $S=2$, it is exactly the conditional likelihood, for any choice of $p_{\lambda}(\lambda)$.

Example 4.2 (Gamma with common shape parameter)
Let $Y_{ij}, i = 1,\ldots,q, j = 1,\ldots,m_i$, be independent random variables with Gamma(ψ,λ_i) distribution. For each stratum i,

$$p_{Y_i}(y_i;\psi,\lambda_i) = \exp\{\psi \sum_{j=1}^{m_i} \log y_{ij} - \lambda_i \sum_{j=1}^{m_i} y_{ij} - m_i \log\Gamma(\psi) + m_i\psi\log(\lambda_i)\} \prod_{j=1}^{m_i} y_{ij}^{-1}$$

If we set the natural observation $\left(\sum_{j=1}^{m_i} \log y_{ij}, \sum_{j=1}^{m_i} y_{ij}\right) = (t_i, s_i)$ we obtain (see Pace and Salvan, 1997, Example 5.13),

$$M(\psi, \lambda_i) = \frac{\Gamma(\psi)^{m_i}}{\lambda_i^{m_i\psi}},$$

and

$$M_{s_i}(\psi) = \frac{s_i^{m_i\psi}\Gamma(\psi)^{m_i}}{\Gamma(m_i\psi)}.$$

Substituting these quantities in (4.4), the weight function is

$$\begin{aligned} g(\lambda_i; \psi) &= \frac{\Gamma(\psi)^{m_i}}{\lambda_i^{m_i\psi}} \left(\frac{s_i^{m_i\psi}\Gamma(\psi)^{m_i}}{\Gamma(m_i\psi)} \right)^{-1} p_{\lambda_i}(\lambda_i) \\ &= \frac{\Gamma(m_i\psi)}{\lambda_i^{m_i\psi} s_i^{m_i\psi}} p_{\lambda_i}(\lambda_i), \end{aligned}$$

and the corresponding integrated likelihood is proportional to the conditional likelihood,

$$\begin{aligned} L_I(\psi) =& \prod_{i=1}^{q} \int_{\mathbb{R}^+} \frac{e^{\psi t_i}\lambda_i^{m_i\psi}}{\Gamma(\psi)^{m_i}} \frac{s_i^{\lambda_i}\Gamma(m_i\psi)}{\lambda_i^{m_i\psi} s_i^{m_i\psi}} p_{\lambda_i}(\lambda_i) d\lambda_i \\ \propto& \prod_{i=1}^{q} \frac{e^{\psi t_i}\Gamma(m_i\psi)}{s_i^{m_i\psi}\Gamma(\psi)^{m_i}} = L_C(\psi). \end{aligned}$$

4.2.1 Approximation with the modified profile likelihood

A major drawback for using (4.4) is the potential difficulty of obtaining the conditional moment generating function. To overcome this issue, we can use an approximation for it, as given by Pace and Salvan (1992),

$$\tilde{M}_s(\psi) = \exp\{\tilde{K}_s(\psi)\} = \exp\{K(\psi, \hat{\lambda}_\psi) - \hat{\lambda}_\psi s - \frac{1}{2}\log|K_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|\}. \quad (4.5)$$

If we substitute this expression for $M_s(\psi)$ in the expression of $L_I(\psi)$, we obtain

$$\begin{aligned}
L_I(\psi) &= \int_\Lambda L(\psi,\lambda)\frac{M(\psi,\lambda)}{\tilde{M}_s(\psi)}g(\lambda)d\lambda \qquad (4.6)\\
&= \int_\Lambda e^{\lambda s+\psi t-K(\psi,\lambda)}\frac{e^{K(\psi,\lambda)}e^{\frac{1}{2}\log|K_{\lambda\lambda}(\psi,\hat{\lambda}_\psi)|}}{e^{K(\psi,\hat{\lambda}_\psi)}e^{-\hat{\lambda}_\psi s}}g(\lambda)d\lambda\\
&= \int_\Lambda e^{\hat{\lambda}_\psi s+\psi t-K(\psi,\hat{\lambda}_\psi)}e^{\frac{1}{2}\log|K_{\lambda\lambda}(\psi,\hat{\lambda}_\psi)|}e^{\lambda s}g(\lambda)d\lambda\\
&= L_P(\psi)|K_{\lambda\lambda}(\psi,\hat{\lambda}_\psi)|^{1/2}\int_\Lambda e^{\lambda s}g(\lambda)d\lambda\\
&\propto L_{MP}(\psi),
\end{aligned}$$

where $L_{MP}(\psi)$ is the modified profile likelihood (Barndorff-Nielsen, 1983), which coincides with the double saddlepoint approximation of the conditional likelihood based on the conditional model for T given $S = s$.

Example 4.2 (continued) (Gamma with common shape parameter)
The likelihood is

$$L(\psi,\lambda) = \prod_{i=1}^{q}\frac{\lambda_i^{m_i\psi}}{\Gamma(\psi)^{m_i}}\left(\prod_{j=1}^{m_i}y_{ij}\right)^{\psi-1}\exp\{-\lambda_i\sum_{j=1}^{m_i}y_{ij}\}.$$

While we leave unchanged the joint moment generating function, we substitute $M_s(\psi)$ with its approximation (4.5),

$$\tilde{K}_s(\psi) = K(\psi,\hat{\lambda}_{\psi i}) - \hat{\lambda}_{\psi i}s - \frac{1}{2}\log|K_{\lambda_i\lambda_i}(\psi,\hat{\lambda}_{\psi i})|, \qquad (4.7)$$

where

$$\hat{\lambda}_{\psi i} = \frac{m_i\psi}{\sum_{j=1}^{m_i}y_{ij}}.$$

Computing each term of (4.7),

$$\begin{aligned}
K(\psi,\hat{\lambda}_{\psi i}) &= m_i\log\Gamma(\psi) - m_i\psi\log\left(\frac{m_i\psi}{\sum_{j=1}^{m_i}y_{ij}}\right)\\
K_{\lambda_i\lambda_i}(\psi,\hat{\lambda}_{\psi i}) &= \frac{\partial^2K_{\lambda_i\lambda_i}(\psi,\lambda_i)}{\partial\lambda_i^2}\Big|_{\lambda_i=\hat{\lambda}_{\psi i}} = \frac{\sum_{j=1}^{m_i}y_{ij}}{m_i\psi},
\end{aligned}$$

we obtain the expression for the approximate moment generating function,

$$\begin{aligned}\tilde{M}_s(\psi) =& \exp\{\tilde{K}_s(\psi)\} = \exp\left\{ m_i \log\Gamma(\psi) - m_i\psi \log\left(\frac{m_i\psi}{\sum_{j=1}^{m_i} y_{ij}}\right)\right.\\ &\left. + \frac{m_i\psi}{\sum_{j=1}^{m_i} y_{ij}} \sum_{j=1}^{m_i} y_{ij} - \frac{1}{2}\log\left|\frac{(\sum_{j=1}^{m_i} y_{ij})^2}{m_i\psi}\right|\right\}\\ =& \frac{\Gamma(\psi)^{m_i}\exp\{m_i\psi\}}{\left(\frac{m_i\psi}{\sum_{j=1}^{m_i} y_{ij}}\right)^{m_i\psi}\left|\frac{(\sum_{j=1}^{m_i} y_{ij})^2}{m_i\psi}\right|^{1/2}}.\end{aligned}$$

With both $M(\psi, \lambda)$ and $\tilde{M}_s(\psi)$, we can compute (4.6),

$$\begin{aligned}L_I(\psi) =& \prod_{i=1}^{q}\int_\Lambda \frac{\lambda_i^{m_i\psi}}{\Gamma(\psi)^{m_i}}\left(\prod_{j=1}^{m_i} y_{ij}\right)^{\psi-1}\\ & \exp\{-\lambda_i \sum_{j=1}^{m_i} y_{ij}\}\frac{\Gamma(\psi)^{m_i}}{\lambda_i^{m_i\psi}}\frac{\left(\frac{m_i\psi}{\sum_{j=1}^{m_i} y_{ij}}\right)^{m_i\psi}\left|\frac{(\sum_{j=1}^{m_i} y_{ij})^2}{m_i\psi}\right|^{1/2}}{\Gamma(\psi)^{m_i}\exp\{m_i\psi\}} g(\lambda)d\lambda\\ =& \prod_{i=1}^{q}\frac{(m_i\psi)^{m_i\psi-\frac{1}{2}}\exp\{-m_i\psi\}\left(\prod_{j=1}^{m_i} y_{ij}\right)^{\psi}}{\Gamma(\psi)^{m_i}\left(\sum_{j=1}^{m_i} y_{ij}\right)^{m_i\psi}}\int_\Lambda \frac{\exp\{-\lambda_i\sum_{j=1}^{m_i} y_{ij}\}\sum_{j=1}^{m_i} y_{ij}}{\prod_{j=1}^{m_i} y_{ij}} g(\lambda)d\lambda\\ \propto& \prod_{i=1}^{q}\frac{(m_i\psi)^{m_i\psi-\frac{1}{2}}\exp\{-m_i\psi\}\left(\prod_{j=1}^{m_i} y_{ij}\right)^{\psi}}{\Gamma(\psi)^{m_i}\left(\sum_{j=1}^{m_i} y_{ij}\right)^{m_i\psi}} = L_{MP}(\psi),\end{aligned}$$

where the expression of $L_{MP}(\psi)$ can also be found in Sartori (2003, Example 2).

4.2.2 Remarks

In this section we have suggested a weight function which leads to the equivalence between conditional and integrated likelihood, forcing the integral (4.2) to be independent of ψ. We can see this equivalence directly, as done in (4.6) with L_I versus L_{MP}. Indeed

$$L_C(\psi) = \exp\{t\psi - K_s(\psi)\}$$

corresponds to an integrated likelihood with weight function (4.4)

$$\begin{aligned} L_I(\psi) &= \int \exp\{t\psi + s\lambda - K(\psi,\lambda)\}g(\lambda;\psi)d\lambda \\ &= \int \exp\{t\psi + s\lambda - K(\psi,\lambda)\}\exp\{K(\psi,\lambda) - K_s(\psi)\}d\lambda \\ &= \exp\{t\psi - K_s(\psi)\}\int \exp\{s\lambda\}d\lambda \\ &\propto \exp\{t\psi - K_s(\psi)\}. \end{aligned}$$

A feature of the proposed weight function is that it depends on the data. While this is not allowed in a fully Bayesian inference, it is already exploited in literature in different frameworks (see, for example Wasserman, 2000; Severini, 2007).

4.3 Weight function that depends on a hyperparameter

Consider now a weight function which depends on a hyperparameter ξ. To better appreciate the advantage of this kind of weight function, we focus on the stratified data case. Let $y_i,\ i = 1,\dots,q$ be a realization of an $m-$dimensional random variable Y_i with density $p_{Y_i}(y_i;\psi,\lambda_i)$. The parameter of interest ψ is common to all strata, while the nuisance parameter λ_i is stratum-specific. Suppose that $Y_1,\dots,Y_q$ are independent and let s_i be a sufficient statistic for λ_i with finite support $\mathcal{S}$. Denote with c the cardinality of $\mathcal{S}$, i.e. set $c = |\mathcal{S}|$.

Under the same conditions which give factorization (4.1) in Section 4.2, with the addition of the hyperparameter, and taking into consideration the stratified nature of the data, it follows that the integrated log-likelihood is

$$l_I(\psi,\xi) = l_C(\psi) + \sum_{i=1}^{q}\log\int_\Lambda p_{S_i}(s_i;\psi,\lambda_i)g(\lambda_i;\xi)d\lambda_i. \tag{4.8}$$

It depends on both the parameter of interest and the hyperparameter. The latter, however, summarizes the information about the nuisance parameters contained in each stratum, and, therefore, its dimension does not increase with the sample size. In particular, this allows us to make inference on ψ through the traditional profile log-likelihood for ψ, without encountering the incidental parameters problem. To this end, we have to solve the constrained likelihood equation

$$\frac{\partial l_I(\psi,\xi)}{\partial \xi} = \sum_{i=1}^{q} \frac{\int_\Lambda p_{S_i}(s_i;\psi,\lambda_i)\frac{\partial}{\partial\xi}g(\lambda_i;\xi)d\lambda_i}{\int_\Lambda p_{S_i}(s_i;\psi,\lambda_i)g(\lambda_i;\xi)d\lambda_i} = 0.$$

Plugging-in the solution, namely $\hat{\xi}_\psi$, into (4.8), we obtain the *profile integrated log-likelihood*,

$$l_{IP}(\psi) = l_I(\psi,\hat{\xi}_\psi) = l_C(\psi) + \sum_{i=1}^{q} \log \int_\Lambda p_{S_i}(s_i;\psi,\lambda_i)g(\lambda_i;\hat{\xi}_\psi)d\lambda_i. \qquad (4.9)$$

We can notice that, in the random effects model framework, the profile integrated log-likelihood corresponds to the usual log-likelihood for the fixed parameter of interest. Moreover, there is a relation between our method and the empirical Bayes technique, since both proceed by summarizing the uncertainty due to incidental nuisance parameters into a unique hyperparameter estimated through the data.

The idea to achieve the equivalence between profile integrated and conditional likelihood is similar to the one presented in the previous section. Indeed, the two approaches, conditioning and integration, produce equivalent log-likelihoods for ψ when (4.2) does not depend on ψ, and, in this case, when the second summand of the right hand side of (4.9) depends only on the data.

First consider the right hand side of (4.8), and, in particular, the second summand. All terms of the sum with the same value of the sufficient statistic provide the same contribution to the log-likelihood. Formally,

$$\sum_{i=1}^{q} \log \int_\Lambda p_{S_i}(s_i;\psi,\lambda_i)g(\lambda_i;\xi)d\lambda_i = \sum_{s\in\mathcal{S}} q_s \log \int_\Lambda p_S(s;\psi,\lambda)g(\lambda;\xi)d\lambda, \qquad (4.10)$$

where q_s is the observed frequency of a given $s \in \mathcal{S}$.

For Rasch models Andersen and Madsen (1977) and Lindsay et al. (1991) noted that (4.10) is the log-likelihood of a multinomial distribution with c possible outcomes, and vector of success probabilities $\pi = \pi(\psi,\xi) = [\pi_s]$, $s \in \mathcal{S}$, where

$$\pi_s = \pi_s(\psi,\xi) = \int_\Lambda p_S(s;\psi,\lambda)g(\lambda;\xi)d\lambda. \qquad (4.11)$$

In fact, this is true for any model, provided that $g(\lambda_i;\xi)$ is a proper density. For (4.10) to be a multinomial log-likelihood, we must have $\sum_{s\in\mathcal{S}}\pi_s = 1$. Given that

$$\begin{aligned}\sum_{s\in\mathcal{S}}\int_\Lambda p_S(s;\psi,\lambda)g(\lambda;\xi)d\lambda &= \int_\Lambda\sum_{s\in\mathcal{S}}p_S(s;\psi,\lambda)g(\lambda;\xi)d\lambda\\ &= \int_\Lambda g(\lambda;\xi)d\lambda,\end{aligned}$$

since $p_S(s;\psi,\lambda)$ is a probability function, it is clear that $\sum_{s\in\mathcal{S}}\pi_s = 1$ only if $\int_\Lambda g(\lambda;\xi)d\lambda = 1$.

Consider now the maximization step which leads to the constrained estimate $\hat{\xi}_\psi$. It is well known that a multinomial likelihood is maximized by

$$\hat{\pi}_s = \frac{q_s}{q},$$

a quantity independent on ψ.

Let Ξ denote the parameter space of the hyperparameter ξ, and Π the $(c-1)$-dimensional codomain consisting of vectors π with $\sum_{s\in\mathcal{S}}\pi_s = 1$, $\pi_s > 0\, s \in \mathcal{S}$. With reference to (4.11), when, for each fixed ψ, $\pi(\psi,\xi)$ is surjective from Ξ to Π, then

$$\int_\Lambda p_S(s;\psi,\lambda)g(\lambda;\hat{\xi}_\psi)d\lambda = \frac{q_s}{q}, \tag{4.12}$$

for any $s\in\mathcal{S}$ and therefore (4.10) is independent on ψ. The probabilities $\pi(\psi,\xi)$ as function of ξ is surjective when, for fixed ψ, every point of Π is an image of at least one point of Ξ.

While from a theoretical point of view it is sufficient to consider a surjective function from Ξ to Π for fixed ψ, it is worth considering $\pi(\psi,\xi)$ as a bijection. This avoids some practical issues; for example, if there are two different points $\xi',\xi''\in\Xi$ such that $\pi(\psi,\xi')=\pi(\psi,\xi'')$, a numerical maximization in ξ for fixed ψ could be troublesome. Moreover, a one-to-one function allows us to think the condition for the equivalence as a reparameterization problem; the equivalence between conditional and profile integrated likelihood holds if, for fixed ψ, we can reparametrize ξ with π, the parameter of the multinomial distribution which leads to (4.10).

Example 4.3 (Trinomial distribution)
As a simple illustration, let $y_i = (y_{i1}, y_{i2}, y_{i3})$, $i = 1, \ldots, q$, be a realization of a trinomial distribution with probabilities p_{i1}, p_{i2}, p_{i3} for the three cells, with $p_{i1} + p_{i2} + p_{i3} = 1$,

$$p_{Y_i}(y_i; p_{i1}, p_{i2}, p_{i3}) = \exp\{y_{i1} \log \frac{p_{i1}}{p_{i3}} + y_{i2} \log \frac{p_{i2}}{p_{i3}} + \log(1 - p_{i1} - p_{i2})\},$$

where y_{ij} is 1 if the observation in stratum i is in cell j and 0 otherwise, $j = 1, 2, 3$.

Assume that $\log(p_{i1}/p_{i2})$ is constant over strata and equal to ψ. Let ψ be the parameter of interest and $\lambda_i = \log(p_{i2}/p_{i3})$ the nuisance parameter for stratum i. With this parameterization,

$$p_{Y_i}(y_i; \psi, \lambda_i) = \exp\{y_{i1}\psi + (y_{i1} + y_{i2})\lambda_i + \log(1 + e^{\psi+\lambda_i} + e^{\lambda_i})\}.$$

Conditioning on the sufficient statistic $s_i = y_{i1} + y_{i2}$, we obtain the conditional log-likelihood

$$l_C(\psi) = \sum_{i=1}^{q_1} y_{i1}\psi - q_1 \log(1 + e^{\psi}),$$

where $q_1 < q$ is the number of the strata where $s_i = 1$.

Since S_i has a Bernulli distribution with success probability $e_i^{\lambda}(1 + e^{\psi})/(1 + e_i^{\lambda}(1 + e^{\psi}))$, relation (4.11) is

$$\pi_1 = \int_{\Lambda} \frac{e_i^{\lambda}(1 + e^{\psi})}{1 + e_i^{\lambda}(1 + e^{\psi})} g(\lambda; \xi) d\lambda. \tag{4.13}$$

If, for instance, we choose as a weight function a Gaussian distribution with unknown mean ξ and variance 1, then, for $\psi = 1$, (4.13) is bijective (see Figure 4.1), and the equivalence between conditional and profile integrated likelihoods holds. We can do the same for other ψ.

This example is useful to understand why Λ must not depend on ψ. Indeed, with a different parameterization, some issues can occur. Specifically, setting $\psi = \pi_{i1}/\pi_{i2}$ and $\lambda_i = \pi_{i2}$, we have

$$p_{Y_i}(y_i; \psi, \lambda_i) = \exp\{y_{i1} \log \psi + (y_{i1} + y_{i2}) \log \frac{\lambda_i}{(1 - \psi\lambda_i - \lambda_i)} + \log(1 - \psi\lambda_i - \lambda_i)\}.$$

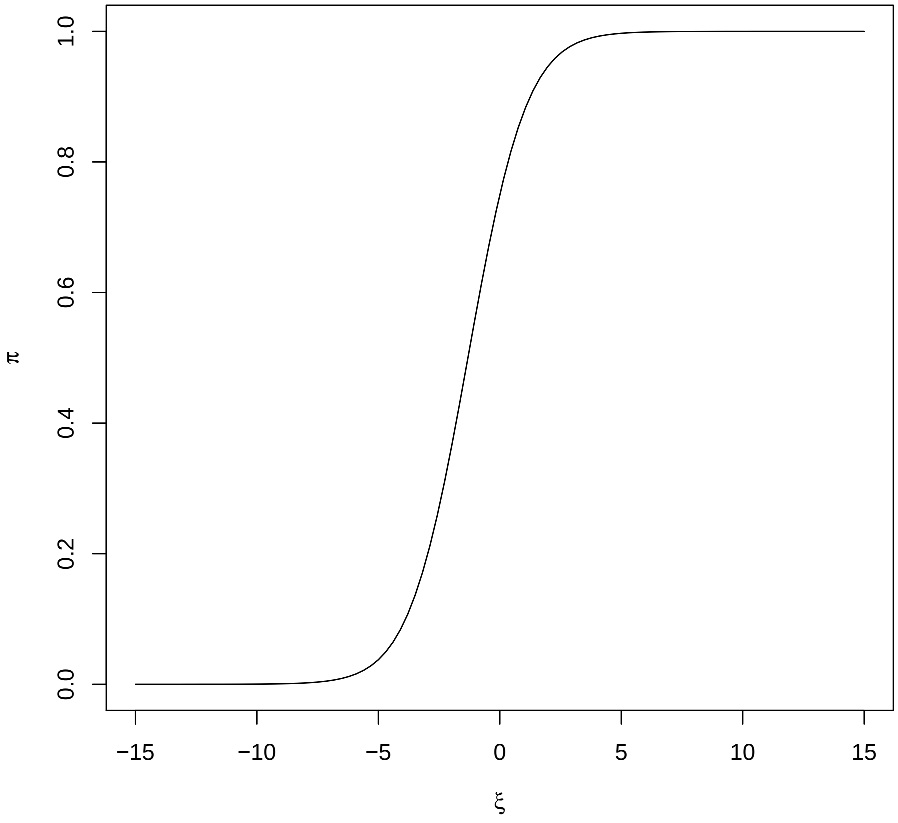

Figure 4.1: Trinomial distribution: function (4.13).

In this case, $1 - \psi\lambda_i - \lambda_i > 0$ implies that $0 < \lambda_i < 1/(1+\psi)$, i.e. the parameter space of λ depends on ψ. Thus, for $\psi > 0$, the range of

$$\int_{\Lambda_\psi} p_{S_i}(s_i;\psi,\lambda_i)g(\lambda_i;\xi)d\lambda_i$$

as ξ varies is not the entire interval $[0,1]$, and if ψ is large the range of this integral is quite small, and could not contain q_{s_i}/q.

Example 4.4 (Matched binary pairs)
Let Y_{i1} and Y_{i2} be two independent Bernoulli random variables with success probabilities p_{i1} and p_{i2}, respectively, $i = 1, \ldots, q$. Here we consider $\psi = \log\{p_{i2}(1-p_{i1})/[p_{i1}(1-p_{i2})]\}$ as the parameter of interest and $\lambda_i = \log\{p_{i1}/(1-p_{i1})\}$ as the nuisance parameters, and denote by $s_i = y_{i1} + y_{i2}$ the marginal totals. The statistic s_i is minimal sufficient for λ_i, for fixed ψ. Conditionally on $S_i = s_i$, Y_{i2} follows a Bernoulli distribution with success probability 0 if $s_i = 0$, 1 if $s_i = 2$ and $e^{-\psi}/(1+e^{-\psi})$ if $s_i = 1$.

If we suppose, without loss of generality, $s_i = 1$ for the first $q_1 \leq q$ strata, the conditional log-likelihood is

$$l_C(\psi) = -\psi\sum_{1=1}^{q_1} y_{i2} - q_1\log(1+e^{-\psi}).$$

To construct a profile integrated log-likelihood, we usually choose as a weight function a location and scale distribution, such as a Gaussian. Consider therefore $g(\lambda_i;\xi) = (1/\tau)\phi\left((\lambda_i-\mu)/\tau\right)$, where $\phi(\cdot)$ is the density of a standard normal distribution, as in Andersen and Madsen (1977), and $\xi = (\mu,\tau)$. The parameter space for ξ is $\Xi = \mathbb{R}\times\mathbb{R}^+$.

Quantity (4.10) becomes

$$\sum_{s=0}^{2} q_s \log\int_\Lambda \frac{e^{\lambda s}\sum_{j=max(0,s-1)}^{min(1,s)} e^{\psi j}}{(1+e^\lambda)(1+e^{\lambda+\psi})}\frac{1}{\sqrt{2\pi\tau}}\exp\{-\frac{1}{2\tau}(\lambda-\mu)^2\}d\lambda. \tag{4.14}$$

This is the log-likelihood of a multinomial distribution with success probabilities

$$\pi_s = \int_\Lambda \frac{e^{\lambda s}\sum_{j=max(0,s-1)}^{min(1,s)} e^{\psi j}}{(1+e^\lambda)(1+e^{\lambda+\psi})}\frac{1}{\sqrt{2\pi\tau}}\exp\{-\frac{1}{2\tau}(\lambda-\mu)^2\}d\lambda, \tag{4.15}$$

with $s = 0, 1, 2$. Straightforward algebra shows that

$$\begin{aligned}\pi_0 + \pi_1 + \pi_2 &= \int_\Lambda \frac{1 + e^\lambda + e^{\lambda+\psi} + e^{2\lambda+\psi}}{(1 + e^\lambda)(1 + e^{\lambda+\psi})} \frac{1}{\sqrt{2\pi\tau}} \exp\{-\frac{1}{2\tau}(\lambda - \mu)^2\} d\lambda \\ &= \int_\Lambda \frac{1}{\sqrt{2\pi\tau}} \exp\{-\frac{1}{2\tau}(\lambda - \mu)^2\} d\lambda = 1.\end{aligned}$$

Unfortunately, the transformation (4.15) is not a proper reparameterization, because there are values of the two-dimensional simplex Π, the parameter space of (π_0, π_1, π_2), which are not images under (4.15) of any point $(\mu, \tau) \in \Xi$. Therefore, the equivalence between conditional and profile integrated likelihood holds only when the unconstrained estimate of (π_0, π_1, π_2), namely $(\frac{q_0}{q}, \frac{q_1}{q}, \frac{q_2}{q})$, lies in the image of Ξ under (4.15) in Π. Denote this set with $\mathcal{I}$.

Let us show this graphically. In this example, fixing $\psi = 1$, we apply (4.15) to a grid of values for μ and τ, with $\mu \in (-50, 50)$ and $\tau \in (0.01, 100)$. The results are presented in Figure 4.2. It shows the images under (4.15) of the considered points on the space Π for π_0 and π_1 (area within the dashed triangle). In particular, small values of τ correspond to the points on the upper contour of the black region, while large values to the lower. Small values of μ correspond to values near the origin, while as μ increases, the corresponding points move far from $(0, 0)$. Values of μ and σ bigger or smaller than the ones considered would not increase in a noteworthy way the areas reported in Figure 4.2. Hence we can state that the space identified by the points in the plot approximates quite well $\mathcal{I}$, and it allows us to understand in broad terms when the equivalence holds and when it does not.

For example, consider $q = 1000$ pairs with $q_0 = 248$, $q_1 = 229$ and $q_2 = 523$. The unconstrained estimate of the multinomial parameter (π_0, π_1, π_2), namely $(0.248, 0.229, 0.523)$ belongs to $\mathcal{I}$, and hence $l_C(1) = l_{IP}(1)$. When we compute the conditional and the profile integrated log-likelihood, however, we obtain two functions which differ from each other by the constant -1022.342. This is the maximum of the multinomial log-likelihood (4.14), namely $\sum_{s\in\{0,1,2\}} q_s \log q_s/q$. The two log-likelihoods, therefore, are completely equivalent, as shown in Figure 4.3.

On the contrary, settings with, for example, $q_0 = 166$, $q_1 = 668$ and $q_2 = 166$ lead to an unconstrained estimate for (π_0, π_1, π_2) which lies outside $\mathcal{I}$.

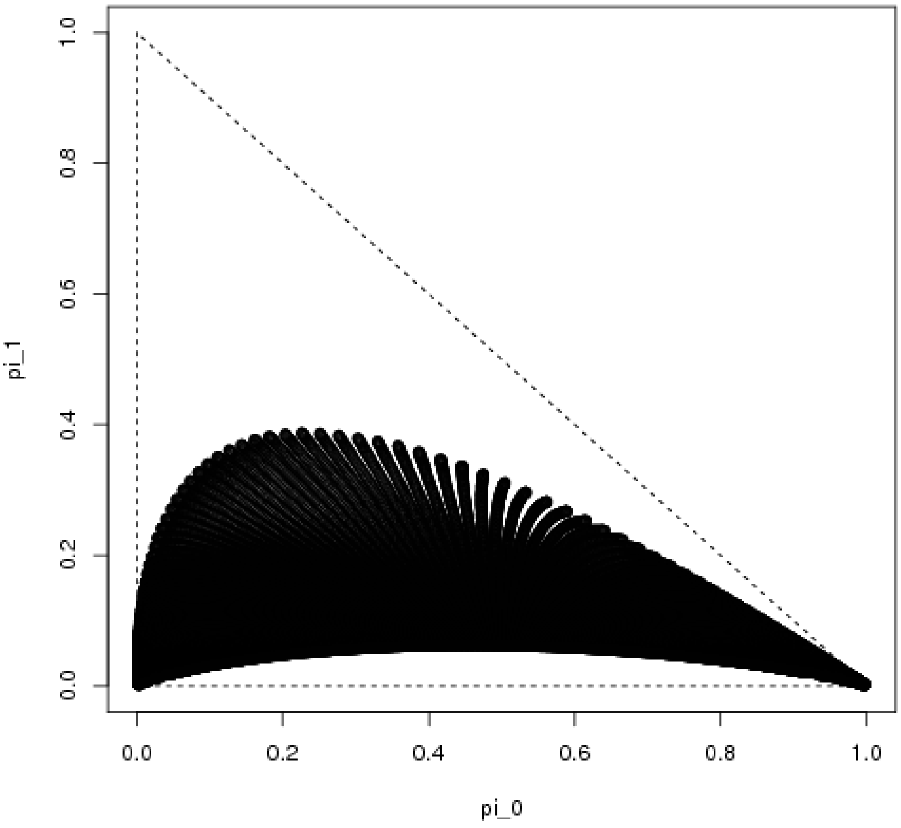

Figure 4.2: Matched binary pairs: images of a grid of (μ, τ) values under (4.15) in Π, when $\psi = 1$.

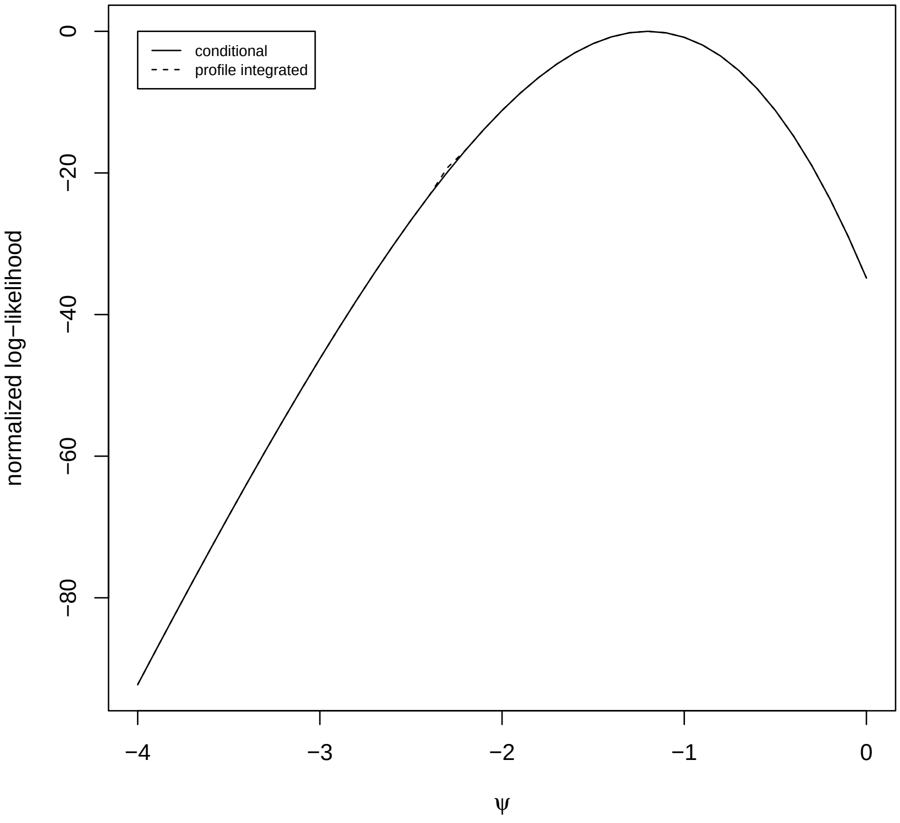

Figure 4.3: Matched binary pairs. Conditional (continuous line) and profile integrated (dashed line) normalized log-likelihood. Data are $q_0 = 248$, $q_1 = 229$, $q_2 = 523$ and $\sum_i yi2 = 576$.

Conditional and profile integrated likelihood, in this case, provide different inferential results, as shown in Figure 4.4.

Example 4.5 (Biallelic genetic marker)
We consider an example taken from Rice (2008, Section 2). It is a case/control study about three unordered categories of exposure, namely dd, dD and DD, of a biallelic genetic marker. Case and control exposures are assumed to be independent in any given pairing, and all pairs are assumed to be independent. The parameter of interest is a genotype-disease association, assessed

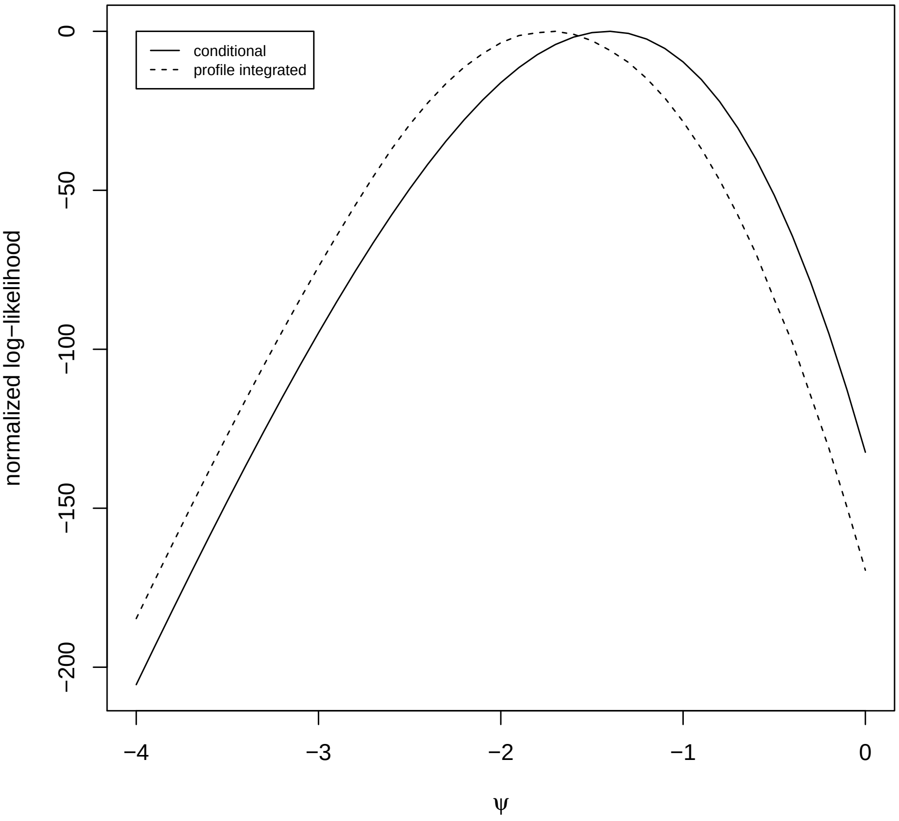

Figure 4.4: Matched binary pairs. Conditional (continuous line) and profile integrated (dashed line) normalized log-likelihood. Data are $q_0 = 166$, $q_1 = 668$, $q_2 = 166$ and $\sum_i yi2 = 197$.

through the odds ratios

$$\psi_1 = \frac{P[\text{case } dD]}{P[\text{case } dd]} \frac{P[\text{control } dd]}{P[\text{control } dD]},$$
$$\psi_2 = \frac{P[\text{case } DD]}{P[\text{case } dd]} \frac{P[\text{control } dd]}{P[\text{control } DD]},$$

that are assumed to be constant across all matched pairs. The reference level of exposure probability in each matched pair i, $i = 1, \ldots, q$, is

Control exposure	dd	dD	DD
Probability	$\frac{1}{1+\lambda_{1i}+\lambda_{2i}}$	$\frac{\lambda_{1i}}{1+\lambda_{1i}+\lambda_{2i}}$	$\frac{\lambda_{2i}}{1+\lambda_{1i}+\lambda_{2i}}$

In this example, the nuisance parameter $\lambda = (\lambda_{1i}, \lambda_{2i})$ has dimension 2 in each stratum, and the sufficient statistic has dimension 2 with $c = 6$ possible values (see Rice, 2008, formula (2)). A natural candidate for the weight function that can lead to equivalence with the conditional likelihood is the bivariate Gaussian distribution with parameter ξ having 5 components (mean and variance-covariance matrix elements).

In this example the conditional likelihood has a simple form (Rice, 2008, formula (3)),

$$L_C(\psi_1, \psi_2) = \left(\frac{\psi_1}{1+\psi_1}\right)^{\#(dd,dD)} \left(\frac{\psi_2}{1+\psi_2}\right)^{\#(dd,DD)} \left(\frac{1}{1+\psi_1}\right)^{\#(dD,dd)}$$
$$\left(\frac{\psi_2}{\psi_1+\psi_2}\right)^{\#(dD,DD)} \left(\frac{1}{1+\psi_2}\right)^{\#(DD,dd)} \left(\frac{\psi_1}{\psi_1+\psi_2}\right)^{\#(DD,dD)},$$

where, for instance, $\#(dd, DD)$ denotes the number of cases with case exposure dd and control exposure DD. The integrated likelihood, instead, is

$$L_I(\psi_1, \psi_2) = \prod_{i=1}^{q} \int_\Lambda \left(\frac{1}{w_i}\right)^{(dd,dd)} \left(\frac{(1+\psi_1)\lambda_{1i}}{w_i}\right)^{(dd,dD)} \left(\frac{(1+\psi_2)\lambda_{2i}}{w_i}\right)^{(dd,DD)}$$
$$\left(\frac{\psi_1\lambda_{1i}^2}{w_i}\right)^{(dD,dD)} \left(\frac{(\psi_1+\psi_2)\lambda_{1i}\lambda_{2i}}{w_i}\right)^{(dD,DD)} \left(\frac{\psi_2\lambda_{2i}}{w_i}\right)^{(DD,DD)}$$
$$g(\lambda_{1i}, \lambda_{2i}; \xi) d(\lambda_{1i}\, \lambda_{2i}),$$

where $w_i = (1 + \lambda_{1i} + \lambda_{2i})(1 + \psi_1\lambda_{1i} + \psi_2\lambda_{2i})$, $g(\lambda_{1i}, \lambda_{2i}; \xi)$ is the bivariate weight function and, for instance, (dd, dd) is 1 if both case and control exposures are dd, 0 otherwise.

As a numerical example, consider $q = 1000$ pairs of variables Y_{i1} and Y_{i2}, with q equal to $(61, 65, 109, 285, 122, 358)$. We used a bivariate Gaussian (5 hyperparameters) and a bivariate Gaussian with identity as variance-covariance matrix (2 hyperparameters) as weight functions. In both cases, integration and optimization steps for the integrated likelihood are performed numerically. In the computation of the profile integrated likelihood, moreover, we have used for the constrained estimates of the hyperparameters the approximation

$$\hat{\xi}_\psi = \hat{\xi} + \left(j_{\xi\xi}(\hat{\psi}, \hat{\xi})\right)^{-1} j_{\xi\psi}(\hat{\psi}, \hat{\xi}) \left(\hat{\psi} - \psi\right), \tag{4.16}$$

see Cox and Wermuth (1990). This causes some approximation errors, as we move away from $\hat{\psi}$, but this is not crucial for our purpose.

In this case, when the weight function is a bivariate Gaussian with 5 unknown parameters, we find a constant difference between the profile integrated and the conditional log-likelihood, equal to $\sum_{s\in\{0,1,2\}} q_s \log q_s/q = -1572.02$. Hence, both log-likelihoods give the same inference. On the contrary, the use of a bivariate normal with identity as variance-covariance matrix leads to a substantially different log-likelihood. Figure 4.5 shows a contour plot of these log-likelihoods. The small discrepancies between the contour lines of normalized conditional and profile integrated log-likelihoods are likely due to the use of approximation (4.16).

Remarks

Rice (2008) shows that there is an equivalence between L_C and L_I if we can find a weight function $g(\lambda;\xi)$ such that $\int_\Lambda p_s(s;\psi,\lambda)g(\lambda;\xi)d\lambda$ is independent of ψ. He reaches this goal substituting ξ with a function $\xi(\psi)$ which is obtained solving a specific moment problem. We have seen that there are cases where $\hat{\xi}_\psi$ satisfies Rice's condition and $\int_\Lambda p_s(s;\psi,\lambda)g(\lambda;\hat{\xi}_\psi)d\lambda$ is independent of ψ. In these situations, L_{IP} and L_C are equivalent. This result shows that, at least in these cases, when we deal with matched binary data an approach based on integration is unable to recover information from concordant pairs.

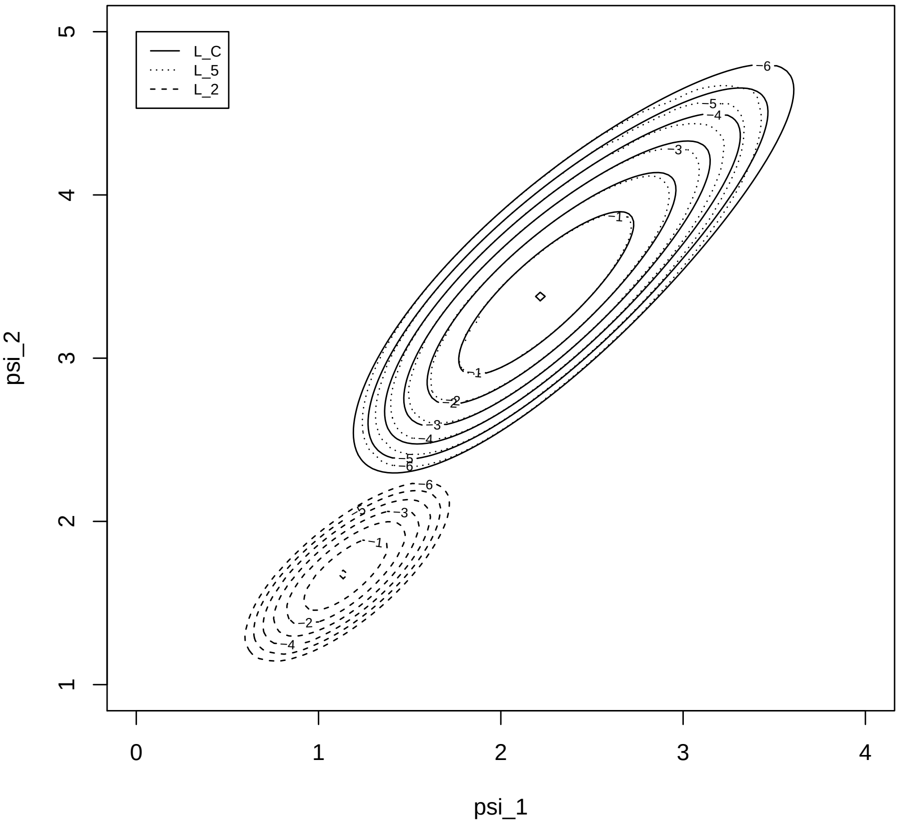

Figure 4.5: Biallelic genetic marker. Normalized log-likelihoods: conditional in continuous line, profile integrated with bivariate Gaussian weight function in dotted line, profile integrated with Gaussian with identity as variance-covariance matrix weight function in dashed line.

Chapter 5

Modified profile pairwise score function

5.1 Introduction

The pairwise likelihood, introduced in Section 2.4, is defined as the product of marginal likelihood contributions, based on the marginal density of two-dimensional components of a vector Y. Let ψ be the parameter of interest and λ the nuisance parameter and let us denote with $p(y_{ir}, y_{is}; \psi, \lambda)$, $r = 1, \ldots, d-1$, $s = r+1, \ldots, d$, $i = 1, \ldots, n$, the marginal density of the pair (Y_{ir}, Y_{is}).

We are interested in making inference about ψ, and in understanding the effect of the presence of the nuisance parameter in the related inferential quantities. In particular, starting from $pl(\psi, \lambda) = \log pL(\psi, \lambda)$ given in (2.8), we want to study the profile pairwise log-likelihood,

$$pl_P(\psi) = pl(\psi, \hat{\lambda}^p_\psi) = \sum_{i=1}^{n} \sum_{s>r} w_{rs} \log p(y_{ir}, y_{is}; \psi, \hat{\lambda}^p_\psi), \tag{5.1}$$

and the bias of the corresponding score function. Here $\hat{\lambda}^p_\psi$ is the constrained maximum pairwise likelihood estimate of λ given ψ and w_{rs} some positive weights, that, in the following, we will consider equal to 1. Moreover, let $l^{rs}(\psi, \hat{\lambda}^p_\psi) = \sum_{i=1}^{n} \log p(y_{ir}, y_{is}; \psi, \hat{\lambda}^p_\psi)$ denotes the profile log-likelihood related to the generic pair (y_r, y_s) and a suitable subscript its derivative with respect to the parameter, for example $l^{rs}_\psi(\psi, \hat{\lambda}^p_\psi) = \frac{\partial}{\partial \psi} l^{rs}(\psi, \hat{\lambda}^p_\psi)$.

As for usual log-likelihood, the profile pairwise score is typically biased. In order to quantify the order of such bias, we consider a two-index asymptotics setting, where we allow both the sample size, n, and the dimension d of the vector Y, to grow to infinity. In this setting, we will see that

$$E[pl_{P\psi}(\psi)] = E\left[\frac{\partial pl(\psi, \hat{\lambda}^p_\psi)}{\partial \psi}\right] = O(d^2). \tag{5.2}$$

A seminal paper that studies pairwise likelihood quantities is Cox and Reid (2004). In that paper, the asymptotic properties of the maximum likelihood estimator of a single parameter θ have been analysed, through the expansion of the estimating equation $pl_\theta(\theta) = 0$.

Cox and Reid (2004) pointed out that, for fixed n, the variance of the estimator is $O(1)$ in d, i.e. the estimating equation based on pairwise likelihood does not usually lead to a consistent estimator. When n grows to infinity, instead, since the observations are independent, the usual asymptotic theory works. This should be true even when d is allowed to grow to infinity.

In this chapter we report a preliminary analysis done about the profile pairwise score bias and a possible modification to the pairwise profile estimating equation useful to reduce it. In Section 5.2 we study the profile pairwise score bias in a two-index asymptotics setting, following the outline of the analysis done by Adimari and Ventura (2002) and by Severini (2002) in a regular one-index asymptotic setting. Moreover, in Section 5.3, we show that the correction proposed by the latter holds also in the pairwise likelihood framework. A brief discussion is presented in Section 5.4.

5.2 Expansion of the pairwise score bias

Let us consider, for notational simplicity, λ scalar, although all the results are valid also in the multidimensional case. From the expansion of the pairwise score function around the true parameter value,

$$pl_{\psi}(\psi, \hat{\lambda}^{p}_{\psi}) = pl_{\psi}(\psi, \lambda) + pl_{\psi\lambda}(\psi, \lambda)Z + \frac{1}{2}pl_{\psi\lambda\lambda}(\psi, \lambda)Z^2 + \dots, \tag{5.3}$$

we can find an approximation of the profile pairwise score bias,

$$\begin{aligned} E[pl_{\psi}(\psi, \hat{\lambda}_{\psi})] &= E[pl_{\psi\lambda}(\psi, \lambda)Z] + \frac{1}{2}E[pl_{\psi\lambda\lambda}(\psi, \lambda)Z^2] + O(d^2/n) \\ &= E[\nu_{\psi\lambda}Z] + E[(pl_{\psi\lambda}(\psi, \lambda) - \nu_{\psi\lambda})Z] + \frac{1}{2}E[\nu_{\psi\lambda\lambda}Z^2] + O(d^2/n). \end{aligned} \tag{5.4}$$

Here we denote by Z the quantity $(\hat{\lambda}^{p}_{\psi} - \lambda)$ and with ν and the suitable subscripts the expected values of the derivatives of the pairwise likelihood quantities; for example we have $\nu_{\psi\lambda} = E[pl_{\psi\lambda}(\psi, \lambda)]$ and $\nu_{\psi,\psi\lambda} = E[pl_{\psi}(\psi, \lambda)\, pl_{\psi\lambda}(\psi, \lambda)]$. Consider now the asymptotic order of these quantities. The order of quatities $\nu_{\psi\lambda}$, $\nu_{\psi\lambda\lambda}$, $\nu_{\lambda\lambda}$, etc. is $O(nd^2)$, since they are sums of $d(d-1)$ terms of order $O(n)$. For example

$$\nu_{\psi\lambda} = E[\sum_{s>r} l^{rs}_{\psi\lambda}(\psi, \lambda)] = \sum_{s>r} E[l^{rs}_{\psi\lambda}(\psi, \lambda)].$$

Terms $\nu_{\psi,\psi\lambda}$, $\nu_{\psi,\lambda}$, etc., instead, are of order $O(nd^4)$. For example,

$$\begin{aligned}
&\nu_{\psi,\psi\lambda} = E[\sum_{s>r} l^{rs}_{\psi}(\psi,\lambda) \sum_{s>r} l^{rs}_{\psi\lambda}(\psi,\lambda)] \\
&= \frac{d(d-1)}{2} E[l^{rs}_{\psi}(\psi,\lambda) l^{rs}_{\psi\lambda}(\psi,\lambda)] + d(d-1)(d-2) E[l^{rs}_{\psi}(\psi,\lambda) l^{st}_{\psi\lambda}(\psi,\lambda)] \\
&+ \frac{d(d-1)(d-2)(d-3)}{4} E[l^{rs}_{\psi}(\psi,\lambda) l^{tu}_{\psi\lambda}(\psi,\lambda)].
\end{aligned}$$

Going back to expansion (5.4), in order to proceed with the approach of Severini (2002), we need an expansion for Z. Starting from the estimating equation $pl_{\lambda}(\psi, \hat{\lambda}^p_{\psi}) = 0$, expanding it around the true value and inverting the resulting expression into an asymptotic expansion for Z (as done by Adimari and Ventura (2002) for usual likelihood, see also Pace and Salvan (1997, Section 9.4.1)), we obtain

$$\begin{aligned}
Z = &- \frac{pl_{\lambda}(\psi,\lambda)}{\nu_{\lambda\lambda}} - \frac{1}{2} \frac{pl_{\lambda}(\psi,\lambda) pl_{\lambda}(\psi,\lambda)}{\nu^3_{\lambda\lambda}} \nu_{\lambda\lambda\lambda} + \frac{pl_{\lambda\lambda}(\psi,\lambda) - \nu_{\lambda\lambda}}{\nu^2_{\lambda\lambda}} pl_{\lambda}(\psi,\lambda) \\
&+ \frac{1}{6} \left(\frac{\nu_{\lambda\lambda\lambda\lambda}}{\nu_{\lambda\lambda}} - 3 \frac{\nu^2_{\lambda\lambda\lambda}}{\nu^2_{\lambda\lambda}} \right) \frac{l^3_{\lambda}}{\nu^3_{\lambda\lambda}} + \frac{1}{2} \left(3 \frac{pl_{\lambda\lambda}(\psi,\lambda) - \nu_{\lambda\lambda}}{\nu^2_{\lambda\lambda}} \nu_{\lambda\lambda\lambda} \right. \\
&\left. - \frac{pl_{\lambda\lambda\lambda}(\psi,\lambda) - \nu_{\lambda\lambda\lambda}}{\nu_{\lambda\lambda}} \right) \frac{pl_{\lambda} pl_{\lambda}}{\nu^2_{\lambda\lambda}} - \frac{(pl_{\lambda\lambda}(\psi,\lambda) - \nu_{\lambda\lambda})^2}{\nu^3_{\lambda\lambda}} pl_{\lambda} + \ldots
\end{aligned}$$

In particular, the first summand is of order $O_p(1/\sqrt{n})$, the second and the third $O_p(1/n)$, and the following $O_p(1/n^{3/2})$. Indeed, $pl_{\lambda}(\psi,\lambda)$ has mean 0 and variance $O_p(nd^4)$ (see also Cox and Reid, 2004) and $pl_{\lambda\lambda}(\psi,\lambda) = \sum_{s>r} l^{rs}_{\lambda\lambda}(\psi,\lambda) = O_p(nd^2)$. Moreover, both $(pl_{\lambda\lambda} - \nu_{\lambda\lambda})$ and $(pl_{\lambda\lambda\lambda} - \nu_{\lambda\lambda\lambda})$ are $O(\sqrt{n}d^2)$, since they have mean 0 and variance $O_p(nd^4)$,

$$\begin{aligned}
&Var[pl_{\lambda\lambda}(\psi,\lambda) = E[(\sum_{s>r} l^{rs}_{\lambda\lambda}(\psi,\lambda))^2] \\
&= \frac{d(d-1)}{2} E[(l^{rs}_{\lambda\lambda}(\psi,\lambda))^2] + d(d-1)(d-2) E[l^{rs}_{\lambda\lambda}(\psi,\lambda) l^{st}_{\lambda\lambda}(\psi,\lambda)] \\
&+ \frac{d(d-1)(d-2)(d-3)}{4} E[l^{rs}_{\lambda\lambda}(\psi,\lambda) l^{tu}_{\lambda\lambda}(\psi,\lambda)],
\end{aligned}$$

with $r \neq s \neq t \neq u$.

We are now able to approximate the score bias (5.4) substituting Z with its expansion. The first term is

$$
\begin{aligned}
&E[\nu_{\psi\lambda}Z] = \nu_{\psi\lambda}E[-\frac{pl_{\lambda}(\psi,\lambda)}{\nu_{\lambda\lambda}} - \frac{1}{2}\frac{pl_{\lambda}(\psi,\lambda)pl_{\lambda}(\psi,\lambda)}{\nu_{\lambda\lambda}^{3}}\nu_{\lambda\lambda\lambda} \\
&+ \frac{pl_{\lambda\lambda}(\psi,\lambda) - \nu_{\lambda\lambda}}{\nu_{\lambda\lambda}^{2}}pl_{\lambda}(\psi,\lambda) + O_p(1/n^{3/2})] \\
&= \nu_{\psi\lambda}\left(-\frac{1}{2}\frac{E[pl_{\lambda}(\psi,\lambda)pl_{\lambda}(\psi,\lambda)]}{\nu_{\lambda\lambda}^{3}}\nu_{\lambda\lambda\lambda} + \frac{E[pl_{\lambda\lambda}(\psi,\lambda)pl_{\lambda}(\psi,\lambda)}{\nu_{\lambda\lambda}^{2}} + O(1/n^{2})\right) \\
&= \nu_{\psi\lambda}\left(\frac{\nu_{\lambda\lambda,\lambda}}{\nu_{\lambda\lambda}^{2}} - \frac{1}{2}\frac{\nu_{\lambda,\lambda}}{\nu_{\lambda\lambda}^{3}}\nu_{\lambda\lambda\lambda}\right) + O(d^{2}/n),
\end{aligned}
$$

the second term is

$$
\begin{aligned}
&E[(pl_{\psi\lambda}(\psi,\lambda) - \nu_{\psi\lambda})Z] = -\frac{E[pl_{\psi\lambda}(\psi,\lambda)pl_{\lambda}(\psi,\lambda)]}{\nu_{\lambda\lambda}} - \frac{\nu_{\psi\lambda}E[pl_{\lambda}(\psi,\lambda)]}{\nu_{\lambda\lambda}} \\
&- \frac{1}{2}\frac{E[(pl_{\psi\lambda}(\psi,\lambda) - \nu_{\psi\lambda})pl_{\lambda}(\psi,\lambda)pl_{\lambda}(\psi,\lambda)]}{\nu_{\lambda\lambda}^{3}}\nu_{\lambda\lambda\lambda} \\
&+ \frac{E[(pl_{\lambda\lambda}(\psi,\lambda) - \nu_{\lambda\lambda})^{2}pl_{\lambda}(\psi,\lambda)]}{\nu_{\lambda\lambda}^{2}} + \dots \\
&= -\frac{\nu_{\psi\lambda,\lambda}}{\nu_{\lambda\lambda}} + O(d^{2}/n),
\end{aligned}
$$

and the third term is

$$
\begin{aligned}
&E[\nu_{\psi\lambda\lambda}Z^{2}] = \nu_{\psi\lambda\lambda}\left(\frac{E[pl_{\lambda}(\psi,\lambda)pl_{\lambda}(\psi,\lambda)]}{\nu_{\lambda\lambda}^{2}}\right. \\
&- \frac{1}{2}\frac{E[pl_{\lambda}(\psi,\lambda)pl_{\lambda}(\psi,\lambda)pl_{\lambda}(\psi,\lambda)]}{\nu_{\lambda\lambda}^{4}}\nu_{\lambda\lambda\lambda} \\
&\left.+\frac{E[(pl_{\lambda\lambda}(\psi,\lambda) - \nu_{\lambda\lambda})pl_{\lambda}(\psi,\lambda)pl_{\lambda}(\psi,\lambda)]}{\nu_{\lambda\lambda}^{3}} + \dots\right) \\
&= \frac{\nu_{\lambda,\lambda}}{\nu_{\lambda\lambda}^{2}}\nu_{\psi\lambda\lambda} + O(d^{2}/n).
\end{aligned}
$$

In these computations, we use the fact that $E[(pl_{\psi\lambda} - \nu_{\psi\lambda})pl_{\lambda}pl_{\lambda}]$ can be rewritten as

$$
\sum_{s>r}\sum_{u>t}\sum_{w>v} E[H_{\psi\lambda}^{rs}(\psi,\lambda)l_{\lambda}^{tu}(\psi,\lambda)l_{\lambda}^{vw}(\psi,\lambda)],
$$

that is a sum of components of order $O(n)$. The number of such components is of order $O(d^{6})$, and therefore $E[(pl_{\psi\lambda} - \nu_{\psi\lambda})pl_{\lambda}pl_{\lambda}]$ is a quantity of order $O(nd^{6})$. Here we denote by $H_{\psi\lambda\lambda}^{rs}$ the quantity $(l_{\psi\lambda\lambda}^{rs} - \nu_{\psi\lambda\lambda}^{rs})$, where $\nu_{\psi\lambda\lambda}^{rs} =$

$E[l^{rs}_{\psi\lambda\lambda}]$. Moreover, also $E[pl_\lambda pl_\lambda pl_\lambda]$ is of order $O(nd^6)$, since it is equal to

$$\sum_{s>r}\sum_{u>t}\sum_{w>v} E[l^{rs}_\lambda(\psi,\lambda)l^{tu}_\lambda(\psi,\lambda)l^{vw}_\lambda(\psi,\lambda)].$$

Combining these results, the profile pairwise score bias is therefore

$$\frac{\nu_{\psi\lambda}}{\nu^2_{\lambda\lambda}}\left(\nu_{\lambda\lambda,\lambda}-\frac{1}{2}\frac{\nu_{\lambda,\lambda}}{\nu_{\lambda\lambda}}\nu_{\lambda\lambda\lambda}\right)-\frac{\nu_{\psi\lambda,\lambda}}{\nu_{\lambda\lambda}}+\frac{1}{2}\frac{\nu_{\lambda,\lambda}\nu_{\psi\lambda\lambda}}{\nu^2_{\lambda\lambda}}+O(d^2/n). \tag{5.5}$$

where the leading term is of order $O(d^2)$.

5.3 Correction to the profile pairwise likelihood estimating equation

Starting from a generic estimating function of the form $(g_\psi(\psi,\lambda), g_\lambda(\psi,\lambda))$, Severini (2002) proposed a modification to general estimating equations which reduces the order of the bias of a general profile score $g_\psi(\psi,\tilde\lambda_\psi)$ from $O(1)$ to $O(1/n)$ in the standard one-index asymptotic setting. Here $\tilde\lambda_\psi$ denotes the solution of $g_\lambda(\psi,\lambda)=0$. This modification, namely

$$\frac{1}{2}\text{tr}\{Dg_{\lambda\lambda}(\psi,\tilde\lambda_\psi)^{-1}\frac{\partial}{\partial\psi}g_{\lambda\lambda}(\psi,\tilde\lambda_\psi)\}-\text{tr}\{DI(\psi,\tilde\lambda_\psi;\tilde\psi,\tilde\lambda)^{-1}\frac{\partial}{\partial\psi}I(\psi,\tilde\lambda_\psi;\tilde\psi,\tilde\lambda)\},$$

with $D=\{-g_{\lambda\lambda}(\psi,\hat\lambda_\psi)\}^{-1}I(\psi,\tilde\lambda_\psi;\tilde\psi,\tilde\lambda)$, is based on the quantities

$$g_{\lambda\lambda}(\psi,\tilde\lambda_\psi)=\frac{\partial}{\partial\lambda}g_\lambda(\psi,\lambda)|_{\lambda=\tilde\lambda_\psi},$$

and

$$I(\psi,\lambda;\tilde\psi,\tilde\lambda)=E_{(\psi_0,\lambda_0)}[g_\lambda(\psi,\lambda)g_\lambda(\psi_0,\lambda_0)]|_{(\psi_0,\lambda_0)=(\tilde\psi,\tilde\lambda)}$$

computed in $\lambda=\tilde\lambda_\psi$. Here $\tilde\psi$ is the solution of $g_\psi(\psi,\tilde\lambda_\psi)=0$ and $\tilde\lambda=\tilde\lambda_{\tilde\psi}$.

Let us consider the corresponding quantities in the pairwise likelihood setting, $pl_{\lambda\lambda}(\psi,\hat\lambda^p_\psi)$ and $E_{(\hat\psi^p,\hat\lambda^p)}[pl_\lambda(\psi,\hat\lambda^p_\psi)pl_\lambda(\hat\psi^p,\hat\lambda^p)]$ respectively. Here $(\hat\psi^p,\hat\lambda^p)$ denotes the maximum pairwise likelihood estimate for (ψ,λ). Consider first

$$[-pl_{\lambda\lambda}(\psi,\hat\lambda_\psi)^p]^{-1}\frac{\partial[-pl_{\lambda\lambda}(\psi,\hat\lambda^p_\psi)]}{\partial\psi}=-\frac{pl_{\lambda\lambda\psi}(\psi,\hat\lambda^p_\psi)+pl_{\lambda\lambda\lambda}(\psi,\hat\lambda^p_\psi)\frac{\partial\hat\lambda^p_\psi}{\partial\psi}}{-pl_{\lambda\lambda}(\psi,\hat\lambda^p_\psi)}.$$

When ψ is the true parameter value,

$$pl_{\lambda\lambda}(\psi, \hat{\lambda}^p_\psi) = \sum_{s>r} l^{rs}_{\lambda\lambda}(\psi, \hat{\lambda}^p_\psi) = \sum_{s>r}(\nu^{rs}_{\lambda\lambda} + O_p(\sqrt{n})) = \nu_{\lambda\lambda} + O_p(d^2\sqrt{n})$$
$$pl_{\lambda\lambda\lambda}(\psi, \hat{\lambda}^p_\psi) = \sum_{s>r} l^{rs}_{\lambda\lambda\lambda}(\psi, \hat{\lambda}^p_\psi) = \sum_{s>r}(\nu^{rs}_{\lambda\lambda\lambda} + O_p(\sqrt{n})) = \nu_{\lambda\lambda\lambda} + O_p(d^2\sqrt{n}),$$

and so on. Moreover, computing the derivative of the estimating equation $pl_\lambda(\psi, \hat{\lambda}^p_\psi) = 0$ with respect to ψ, we find that

$$\frac{\partial \hat{\lambda}^p_\psi}{\partial \psi} = -pl_{\lambda\lambda}(\psi, \hat{\lambda}^p)^{-1} pl_{\lambda\psi}(\psi, \hat{\lambda}^p).$$

Then

$$[-pl_{\lambda\lambda}(\psi, \hat{\lambda}^p_\psi)]^{-1} \frac{\partial[-pl_{\lambda\lambda}(\psi, \hat{\lambda}^p_\psi)]}{\partial \psi} = \frac{\nu_{\psi\lambda\lambda} - \nu_{\lambda\lambda\lambda}\nu_{\psi\lambda}/\nu_{\lambda\lambda}}{\nu_{\lambda\lambda}} \left(1 + O_p\left(\frac{1}{\sqrt{n}}\right)\right).$$

A similar argument holds for

$$pI(\psi, \hat{\lambda}^p_\psi; \hat{\psi}^p, \hat{\lambda}^p) = E_{(\hat{\psi}^p, \hat{\lambda}^p)}[pl_\lambda(\psi, \hat{\lambda}^p_\psi) pl_\lambda(\hat{\psi}^p, \hat{\lambda}^p)].$$

Indeed,

$$[pI(\psi, \hat{\lambda}^p_\psi; \hat{\psi}^p, \hat{\lambda}^p)]^{-1} \frac{\partial pI(\psi, \hat{\lambda}^p_\psi; \hat{\psi}^p, \hat{\lambda}^p)}{\partial \psi} =$$
$$= \frac{E_{(\hat{\psi}^p, \hat{\lambda}^p)}[pl_{\lambda\psi}(\psi, \hat{\lambda}^p_\psi) pl_\lambda(\hat{\psi}^p, \hat{\lambda}^p)] + \frac{\partial \hat{\lambda}^p_\psi}{\partial \psi} E_{(\hat{\psi}^p, \hat{\lambda}^p)}[pl_{\lambda\lambda}(\psi, \hat{\lambda}^p_\psi) pl_\lambda(\hat{\psi}^p, \hat{\lambda}^p)]}{E_{(\hat{\psi}^p, \hat{\lambda}^p)}[pl_\lambda(\psi, \hat{\lambda}^p_\psi) pl_\lambda(\hat{\psi}^p, \hat{\lambda}^p)]},$$

and

$$[pI(\psi, \hat{\lambda}^p_\psi; \hat{\psi}^p, \hat{\lambda}^p)]^{-1} \frac{\partial pI(\psi, \hat{\lambda}^p_\psi; \hat{\psi}^p, \hat{\lambda}^p)}{\partial \psi} =$$
$$= \frac{\nu_{\lambda\psi,\lambda} - \nu_{\lambda\lambda,\lambda}\nu_{\lambda\psi}/\nu_{\lambda\lambda}}{\nu_{\lambda,\lambda}} \left(1 + O_p\left(\frac{1}{\sqrt{n}}\right)\right).$$

Since the term D introduced by Severini (2002) is, in our framework,

$$D = \frac{pI(\hat{\psi}^p, \hat{\lambda}^p; \hat{\psi}^p, \hat{\lambda}^p)}{-pl_{\lambda\lambda}(\hat{\psi}^p, \hat{\lambda}^p)} = -\frac{\nu_{\lambda,\lambda}}{\nu_{\lambda\lambda}} + O_p(d^2/\sqrt{n})),$$

we can state that

$$E\left[D(-pl_{\lambda\lambda}(\psi,\hat{\lambda}^p_\psi))^{-1}\frac{\partial\ -pl_{\lambda\lambda}(\psi,\hat{\lambda}^p_\psi)}{\partial\psi}\right]=-\frac{\nu_{\lambda,\lambda}}{\nu_{\lambda\lambda}}\frac{\nu_{\psi\lambda\lambda}-\nu_{\lambda\lambda\lambda}\frac{\nu_{\psi\lambda}}{\nu_{\lambda\lambda}}}{\nu_{\lambda\lambda}}+O\left(\frac{d^2}{n}\right) \tag{5.6}$$

and

$$E\left[D(pI(\psi,\hat{\lambda}^p_\psi;\hat{\psi}^p,\hat{\lambda}^p))^{-1}\frac{\partial pI(\psi,\hat{\lambda}^p_\psi;\hat{\psi}^p,\hat{\lambda}^p)}{\partial\psi}\right]=-\frac{\nu_{\psi\lambda,\lambda}-\nu_{\lambda\lambda,\lambda}\frac{\nu_{\psi\lambda}}{\nu_{\lambda\lambda}}}{\nu_{\lambda\lambda}}+O\left(\frac{d^2}{n}\right), \tag{5.7}$$

where as usual the order in n drops when taking expectations. Combining (5.6) and (5.7), we obtain the opposite of (5.5). Note that, in order to have this result, we need to take advantage of the symmetry of the sensitivity matrix, i.e. of the fact that $pl_{\psi\lambda}=pl_{\psi\lambda}$ (see Severini, 2002).

Hence,

$$\begin{aligned}E[pl_\psi(\psi,\hat{\lambda}^p_\psi)]+E\Bigg[\frac{1}{2}D(-pl_{\lambda\lambda}(\psi,\hat{\lambda}^p_\psi))^{-1}\frac{\partial\ -pl_{\lambda\lambda}(\psi,\hat{\lambda}^p_\psi)}{\partial\psi}\\ +D(pI(\psi,\hat{\lambda}^p_\psi;\hat{\psi}^p,\hat{\lambda}^p))^{-1}\frac{\partial pI(\psi,\hat{\lambda}^p_\psi;\hat{\psi}^p,\hat{\lambda}^p)}{\partial\psi}\Bigg]=O(d^2/n).\end{aligned}$$

and, therefore, the modification

$$\begin{aligned}&\frac{1}{2}\mathrm{tr}\{Dpl_{\lambda\lambda}(\psi,\hat{\lambda}^p_\psi)^{-1}\frac{\partial}{\partial\psi}pl_{\lambda\lambda}(\psi,\hat{\lambda}^p_\psi)\}\\ &-\mathrm{tr}\{DpI(\psi,\hat{\lambda}^p_\psi;\hat{\psi}^p,\hat{\lambda}^p)^{-1}\frac{\partial}{\partial\psi}pI(\psi,\hat{\lambda}^p_\psi;\hat{\psi}^p,\hat{\lambda}^p)\},\end{aligned} \tag{5.8}$$

obtained adapting the correction proposed by Severini (2002) to the pairwise setting, is useful to reduce the profile pairwise score bias.

The variance of the profile and modified profile pairwise score is $O(nd^4)$. This means that while the relative profile pairwise score bias is

$$\frac{E[pl_{P\psi}(\psi)]}{\sqrt{Var(pl_{P\psi}(\psi))}}=O(1/\sqrt{n}),$$

while the relative score bias using this modification is of order $O(1/\sqrt{n^3})$.

It would be interesting to take advantage of this result to find a correction which acts directly on the profile pairwise log-likelihood, that would approximately recover the first Bartlett identity. When ψ and λ are scalar, this is straightforward. Indeed,

$$\int [\frac{1}{2}D(-pl_{\lambda\lambda}(\psi,\hat{\lambda}^p_\psi))^{-1}\frac{\partial\ -pl_{\lambda\lambda}(\psi,\hat{\lambda}^p_\psi)}{\partial\psi} \\ + D(pI(\psi,\hat{\lambda}^p_\psi;\hat{\psi},\hat{\lambda}^p))^{-1}\frac{\partial pI(\psi,\hat{\lambda}^p_\psi;\hat{\psi}^p,\hat{\lambda}^p)}{\partial\psi}]d\psi = D\log\frac{(-pl_{\lambda\lambda}(\psi,\hat{\lambda}^p_\psi))^{1/2}}{pI(\psi,\hat{\lambda}^p_\psi;\hat{\psi}^p,\hat{\lambda}^p))},$$

and the modified profile pairwise log-likelihood becomes

$$pl_{MP}(\psi) = pl_P(\psi) + D\log\frac{(-pl_{\lambda\lambda}(\psi,\hat{\lambda}^p_\psi))^{1/2}}{pI(\psi,\hat{\lambda}^p_\psi;\hat{\psi}^p,\hat{\lambda}^p)}. \tag{5.9}$$

Unfortunately, when λ is a vector, we have not a close form for the integral. Indeed, in this case the term D is a matrix, and it cannot be taken out from the trace in (5.8), so that explicit integration is not possible. When λ is a vector, therefore, we should apply the modification directly to the estimating equations or compute the integral numerically. Moreover, when ψ is a vector, the integral has not a unique solution.

The study hitherto performed suggests a correction which reduce the score bias only in n. The problem is that, in general, in the pairwise setting the asymptotic theory does not hold in d (Cox and Reid, 2004) for fixed n. In order to recover consistency also for a single observation, Cox and Reid (2004) introduced the condition that $d^{-4}E[pl_\theta(\theta)^2] \to 0$ as $d \to \infty$. They point out that a necessary and sufficient condition to have a consistent estimator and to allow the use of an asymptotic theory in d for n fixed is

$$E[l^{rs}_\theta(\theta)l^{tu}_\theta(\theta)] = 0. \tag{5.10}$$

for $r \neq s \neq t \neq u$. When it holds, the order of the pairwise score variance is $O(d^3)$, and, therefore, the pairwise score function is a quantity of order $O_P(d^{3/2})$. As a consequence, the constrained estimator for the nuisance parameter is $O_p(1/\sqrt{d})$, because

$$(\hat{\lambda}^p_\psi - \lambda) = \frac{pl_\lambda(\psi,\lambda)}{-pl_{\lambda\lambda}(\psi,\lambda)} + O_p(1/d).$$

Moreover, the profile pairwise score can be approximated as

$$pl_\psi(\psi,\hat{\lambda}^p_\psi) = pl_\psi(\psi,\lambda)+pl_{\psi\lambda}(\psi,\lambda)(\hat{\lambda}^p_\psi-\lambda)+\frac{1}{2}pl_{\psi\lambda\lambda}(\psi,\lambda)(\hat{\lambda}^p_\psi-\lambda)^2+O_p(\sqrt{d}).$$

In order to use the regular asymptotic expansions and to follow the previous argument, however, we need more conditions. For example, the order of $(pl_{\psi\lambda}(\psi,\lambda) - \nu_{\psi\lambda})$ is equal to the order of $pl_{\psi\lambda}(\psi,\lambda)$, unless the quantity $E[l^{rs}_{\psi\lambda}(\psi,\lambda) l^{tu}_{\psi}(\psi,\lambda)] = 0$ too. Therefore, with nuisance parameters, the study of the modification for fixed n and increasing d requires additional study.

Example 5.1 (First order autoregression)
Let us consider a normal autoregressive process of order one, as in Example 4.2 of Pace et al. (2011). Using the parameterization given in Davison (2003, Example 6.24),

$$Y_{ir} - \mu = \rho(Y_{ir-1} - \mu) + \epsilon_{ir}$$

where $i = 1,\dots,n$, $r = 2,\dots,d$, and ϵ_{ir} are independent normally distributed with mean 0 and variance τ. This implies that each observation y_i is a realization of a multivariate Gaussian distribution with mean μ and covariance between Y_{ir} and Y_{is} equal to $\tau\rho^{|r-s|}/(1-\rho^2)$, $r,s = 1,\dots,d$.

Consider the partition of the parameter θ in ρ, the component of interest, and $\lambda = (\mu,\tau)$ the nuisance component and focus on a single series, that is $n = 1$. Dropping out the subscript i from the notation, the likelihood is

$$\begin{aligned} l(\theta) = &-\frac{1}{2\tau}\left[\sum_{r=1}^{d}(y_r-\mu)^2 + \rho^2\sum_{r=2}^{d-1}(y_r-\mu)^2 - 2\rho\sum_{r=2}^{d}(y_r-\mu)(y_{r-1}-\mu)\right] \\ &-\frac{d}{2}\log\tau + \frac{1}{2}\log(1-\rho^2). \end{aligned}$$

The pairwise log-likelihood, using only pairs of contiguous components, that is $w_{rs} = 1$ if and only if $|r-s| = 1$, is

$$\begin{aligned} pl(\theta) = &-\frac{1}{2\tau}\left[\sum_{r=2}^{d}(y_r-\mu)^2 + \sum_{r=2}^{d}(y_{r-1}-\mu)^2 - 2\rho\sum_{r=2}^{d}(y_r-\mu)(y_{r-1}-\mu)\right] \\ &-(d-1)\log\tau + \frac{d-1}{2}\log(1-\rho^2). \end{aligned}$$

In this case, the pairwise log-likelihood is in general of order $O(nd)$ and not $O(nd^2)$ as seen in Section 5.3. Solving the score component related to the nuisance parameter, we get the constrained estimates

$$\hat{\mu}^p = \frac{\sum_{r=2}^{d} y_r + \sum_{r=2}^{d} y_{r-1}}{2(d-1)}$$

and

$$\hat{\tau}^p_{\rho\mu} = \frac{\sum_{r=2}^{d}(y_r - mu)^2 + \sum_{r=2}^{d}(y_{r-1} - \mu)^2 - 2\rho \sum_{r=2}^{d}(y_r - \mu)(y_{r-1} - \mu)}{2(d-1)}$$

The information matrix block related to the nuisance parameters $j_{\lambda\lambda}$ has components

$$\begin{aligned}
-pl_{\mu\mu}(\theta) &= \frac{2(1-\rho)(d-1)}{\tau} \\
-pl_{\mu\lambda}(\theta) = pl_{\lambda\mu}(\theta) &= \frac{(1-\rho)[\sum_{r=2}^{d}(y_r - \mu) + \sum_{r=2}^{d}(y_{r-1} - \mu)]}{\tau^2} \\
-pl_{\lambda\lambda}(\theta) &= \frac{\sum_{r=2}^{d}(y_r - \mu)^2 + \sum_{r=2}^{d}(y_{r-1} - \mu)^2 - 2\rho \sum_{r=2}^{d}(y_r - \mu)(y_{r-1} - \mu)}{\tau^3} - \frac{d-1}{\tau^2},
\end{aligned}$$

while the components of $pI(\rho, \hat{\lambda}^p_\rho; \hat{\rho}^p, \hat{\lambda}^p)$ are

$$\begin{aligned}
pI_{\mu\mu}(\hat{\theta}^p_\rho, \hat{\theta}^p) =& \frac{2(1-\rho)}{\hat{\tau}^p_\rho}(\sum_{r=2}^{d}\sum_{s=2}^{d}(\hat{\rho}^p)^{|r-s|} + \sum_{r=2}^{d}\sum_{s=2}^{d}(\hat{\rho}^p)^{|r-s+1|}) \\
pI_{\mu\tau}(\hat{\theta}^p_\rho, \hat{\theta}^p) =& pl_{\tau\mu}(\hat{\theta}^p_\rho, \hat{\theta}^p) = 0 \\
pI_{\tau\tau}(\hat{\theta}^p_\rho, \hat{\theta}^p) =& \frac{1}{(\hat{\tau}^p_\rho)^2(1-(\hat{\rho}^p)^2)^2}\sum_{r=2}^{d}\sum_{s=2}^{d}\Big\{(1-(\hat{\rho}^p)^2)(1-\rho\hat{\rho}^p) + (1+\rho\hat{\rho}^p)(\hat{\rho}^p)^{2|r-s|} \\
& + (\hat{\rho}^p)^{2|r-s+1|} - (\rho + \hat{\rho}^p)((\hat{\rho}^p)^{|r-s|+|r-s+1|} + (\hat{\rho}^p)^{|r-s|+|r-s-1|}) \\
& + (\rho\hat{\rho}^p)(\hat{\rho}^p)^{|r-s+1|+|r-s-1|}\Big\} - \frac{(d-1)^2(1-\rho\hat{\rho}^p)}{(\hat{\tau}^p_\rho)^2(1-(\hat{\rho}^p)^2)}.
\end{aligned}$$

Here $\hat{\theta}_\rho = (\rho, \hat{\mu}, \hat{\tau}^p_{\rho\hat{\mu}})$ and $\hat{\tau}^p_\rho = \hat{\tau}^p_{\rho\hat{\mu}}$.

Simulated data

We simulate $B = 8000$ times an autoregressive model of first order, with mean $\mu = 0$ and variance $\tau = 1$. We compute the empirical bias for estimators based on solution of the profile likelihood score ($\hat{\rho}$), the profile pairwise likelihood score ($\hat{\rho}^p$), the modified profile pairwise likelihood score ($\hat{\rho}^p_M$).

The empirical bias of the estimators are presented in Table 5.1. We can see a small reduction of the empirical bias solving the modified profile score instead of the profile pairwise one. Since we use only the contiguous pairs, the estimate based on full and pairwise likelihood do not differ too much. On the other hand, the estimator based on modified pairwise score function seems to improve, although slightly, also the full maximum likelihood estimator.

Table 5.1: Empirical bias of profile likelihood ($\hat{\rho}$), profile pairwise likelihood ($\hat{\rho}_P$) and modified profile pairwise likelihood ($\hat{\rho}_M$) estimators.

$\rho = 0.25$

	$d = 10$	$d = 30$	$d = 100$
$\hat{E}[\hat{\rho} - \rho]$	-0.082	-0.044	-0.020
$\hat{E}[\hat{\rho}^p - \rho]$	-0.092	-0.045	-0.020
$\hat{E}[\hat{\rho}^p_M - \rho]$	-0.021	-0.021	-0.012

$\rho = 0.5$

	$d = 10$	$d = 30$	$d = 100$
$\hat{E}[\hat{\rho} - \rho]$	-0.214	-0.084	-0.025
$\hat{E}[\hat{\rho}^p - \rho]$	-0.231	-0.087	-0.025
$\hat{E}[\hat{\rho}^p_M - \rho]$	-0.157	-0.064	-0.018

$\rho = 0.75$

	$d = 10$	$d = 30$	$d = 100$
$\hat{E}[\hat{\rho} - \rho]$	-0.314	-0.110	-0.032
$\hat{E}[\hat{\rho}^p - \rho]$	-0.343	-0.117	-0.033
$\hat{E}[\hat{\rho}^p_M - \rho]$	-0.276	-0.100	-0.029

$\rho = 0.9$

	$d = 10$	$d = 30$	$d = 100$
$\hat{E}[\hat{\rho} - \rho]$	-0.384	-0.133	-0.038
$\hat{E}[\hat{\rho}^p - \rho]$	-0.444	-0.146	-0.040
$\hat{E}[\hat{\rho}^p_M - \rho]$	-0.360	-0.126	-0.032

5.4 Remarks

This chapter provided only a first look on the profile pairwise score bias, and some hints about a possible modification to the profile pairwise log-likelihood useful to reduce it. We choose to include it in the thesis in order to show the last efforts done during the PhD program. Anyway, at least an example in two-index asymptotic should be included. Moreover, some efforts should be done in understanding the asymptotic properties of the modification (5.8) when condition (5.10) holds. The example presented, indeed, has been thought in order to understand the behaviour of modification (5.8) as d increases for n fixed. The results seem promising, while additional study is of course required. Finally, as with ordinary likelihood, integration of the pairwise likelihood could be considered as an alternative to profiling.

Bibliography

Adimari, G. and Ventura, L. (2002). Quasi-profile loglikelihoods for unbiased estimating functions. *Annals of the Institute of Statistical Mathematics*, 54, 235–244.

Andersen, E. (1980). *Discrete Statistical Models with Social Science Applications*. North-Holland, Amsterdam.

Andersen, E. and Madsen, M. (1977). Estimating the parameters of the latent population distribution. *Psychometrika*, 42, 357–374.

Barndorff-Nielsen, O. (1980). Conditionality resolutions. *Biometrika*, 67, 293–310.

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

Barndorff-Nielsen, O. (1986). Inference on full or partial parameters based on the standardized signed log-likelihood ratio. *Biometrika*, 73, 307–322.

Barndorff-Nielsen, O. (1994). Adjusted versions of profile likelihood and directed likelihood, and extended likelihood. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56, 125–140.

Barndorff-Nielsen, O. (1995). Stable and invariant adjusted profile likelihood and directed likelihood for curved exponential models. *Biometrika*, 82, 489–499.

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

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

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

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36, 192–236.

Chandler, R. and Bate, S. (2007). Inference for clustered data using the independence loglikelihood. *Biometrika*, 94, 167–183.

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

Cox, D. 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.

Cox, D. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91, 729–737.

Cox, D. and Wermuth, N. (1990). An approximation to maximum likelihood estimates in reduced models. *Biometrika*, 77, 747–761.

Davison, A. (2003). *Statistical Models.* Cambridge University Press, Cambridge.

Davison, A. and Sartori, N. (2008). The Banff challenge: statistical detection of a noisy signal. *Statistical Science*, 23, 354–364.

Follmann, D. (1988). Consistent estimation in the Rasch model based on nonparametric margins. *Psychometrika*, 53, 553–562.

Foutz, R. and Srivastava, R. (1977). The performance of the likelihood ratio test when the model is incorrect. *The Annals of Statistics*, 5, 1183–1194.

Foutz, R. and Srivastava, R. (1978). The asymptotic distribution of the likelihood ratio when the model is incorrect. *Canadian Journal of Statistics*, 6, 273–279.

Fraser, D., Reid, N., and Wong, A. (2004). Inference for bounded parameters. *Physical Review D*, 69, 033002.

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

Kent, J. (1982). Robust properties of likelihood ratio tests. *Biometrika*, 69, 19–27.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 27, 887–906.

Lindsay, B. (1980). Nuisance parameters, mixture models, and the efficiency of partial likelihood estimators. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 296, 639–662.

Lindsay, B. (1985). Using empirical partially Bayes inference for increased efficiency. *The Annals of Statistics*, 13, 914–931.

Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80, 221–239.

Lindsay, B., Clogg, C., and Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, 86, 96–107.

Lindsay, B., Pilla, R., and Basak, P. (2000). Moment-based approximations of distributions using mixtures: theory and applications. *Annals of the Institute of Statistical Mathematics*, 52, 215–230.

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

Mandelkern, M. (2002). Setting confidence intervals for bounded parameters. *Statistical Science*, 17, 149–159.

Maritz, J. (1970). *Empirical Bayes Methods.* Methuen, London.

McCullagh, P. and Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52, 325–344.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* Springer, New York.

Morris, C. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78, 47–55.

Neuhaus, J., Kalbfleisch, J., and Hauck, W. (1994). Conditions for consistent estimation in mixed-effects models for binary matched-pairs data. *Canadian Journal of Statistics*, 22, 139–148.

Neyman, J. and Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16, 1–32.

Pace, L. and Salvan, A. (1992). A note on conditional cumulants in canonical exponential families. *Scandinavian Journal of Statistics*, 19, 185–191.

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference: from a neo-Fisherian perspective.* World Scientific, Singapore.

Pace, L., Salvan, A., and Sartori, N. (2011). Adjusting composite likelihood ratio statistics. *Statistica Sinica*, 21, 129–148.

Pakel, C., Shephard, N., and Sheppard, K. (2011). Nuisance parameters, composite likelihoods and a panel of GARCH models. *Statistica Sinica*, 21, 307–329.

Pfanzagl, J. (1993). Incidental versus random nuisance parameters. *The Annals of Statistics*, 21, 1663–1691.

Pfanzagl, J. and Wefelmeyer, W. (1982). *Contributions to a General Asymptotic Statistical Theory.* Springer, New York.

Pierce, D. and Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54, 701–737.

Pinheiro, J. and Bates, D. (2000). *Mixed-effects Models in S and S-Plus.* Springer, New York.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics*, 16, 356–366.

Rice, K. (2004). Equivalence between conditional and mixture approaches to the Rasch model and matched case-control studies, with applications. *Journal of the American Statistical Association*, 99, 510–523.

Rice, K. (2008). Equivalence between conditional and random-effects likelihoods for pair-matched case-control studies. *Journal of the American Statistical Association*, 103, 385–396.

Robbins, H. (1955). An empirical Bayes approach to statistics. *Proceedings of the Third Berkley Symposium Mathematical Statistics and Probability*, 1, 157–164.

Rotnitzky, A. and Jewell, N. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, 77, 485–497.

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

Sartori, N., Bellio, R., Salvan, A., and Pace, L. (1999). The directed modified profile likelihood with many nuisance parameters. *Biometrika*, 86, 735–742.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2, 110–114.

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

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

Severini, T. (1999). On the relationship between Bayesian and non-Bayesian elimination of nuisance parameters. *Statistica Sinica*, 9, 713–724.

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

Severini, T. (2002). Modified estimating functions. *Biometrika*, 89, 333–343.

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

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

Skovgaard, I. (1989). A review of higher order likelihood inference. *Bulletin of the International Statistical Institute*, 53, 331–351.

Strasser, H. (1996). Asymptotic efficiency of estimates for models with incidental nuisance parameters. *The Annals of Statistics*, 24, 879–901.

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

Sweeting, T. (1995). A framework for Bayesian and likelihood approximations in statistics. *Biometrika*, 82, 1–23.

Sweeting, T. (1996). Approximate Bayesian computation based on signed roots of log-density ratios. *Bayesian Statistics*, 5, 427–444.

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis*, 92, 1–28.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21, 5–42.

Wald, A. (1950). *Statistical Decision Functions.* Wiley, New York.

Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *Journal of the Royal Statistical Society. Series B (Methodological)*, 62, 159–180.

Wood, A. (1989). An F approximation to the distribution of a linear combination of chi-squared variables. *Communications in Statistics - Simulation and Computation*, 18, 1439–1456.

Riccardo De Bin

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 049 827 4123
e-mail: debin@stat.unipd.it

Current Position

Since January 2009; (expected completion: Spring 2012)
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Integrated likelihood for the treatment of the nuisance parameters
Supervisor: Prof. Nicola Sartori
Co-supervisors: Prof. Alessandra Salvan, Prof. Thomas A. Severini.

Research interests

- Asymptotic theory
- Incidental nuisance parameters

Education

October 2005 – October 2007
Master (*laurea specialistica/magistrale*) degree in Statistics and Computer Science .
University of Padova, Faculty of Statistical Sciences

Title of dissertation: “Multimodel inference: a frequentist approach”
Supervisor: Prof. Alessandra Salvan
Final mark: 110 out of 110

October 2002 – July 2005
Bachelor degree (*laurea triennale*) in Statistics and Information Technology.
University of Padova, Faculty of Statistical Sciences
Title of dissertation: “Planning of a call center employers”
Supervisor: Prof. Giovanni Andreatta
Final mark: 110 out of 110 with honours.

Visiting periods

October 2010 – November 2011
Northwestern University,
Evanston (IL), USA.
Supervisor: Prof. Thomas A. Severini

Work experience

June 2008 – December 2008
Marchon®.
Sales forecaster.

February 2008 – June 2008
The North Face®.
Sales Planning Analyst.

Computer skills

- **OS:** DOS, Linux, Windows
- **Programming:** C++, TurboPascal, QuickBasic, Java

- **Others:** R, MySQL, Office, SAS, LaTeX

Language skills

Italian: native; English: good.

Publications

Articles in journals

De Bin, R., Risso, D. (2011). A novel approach to the clustering of microarray data via nonparametric density estimation. *BMC Bioinformatics* **12**:49.

Manuscripts

De Bin, R., On equivalence between conditional and profile integrated likelihood.

De Bin, R., Sartori N., Severini T. A., Integrated likelihood ratio statistic in models with stratum nuisance parameters.

Conference presentations

De Bin, R., Risso, D. (2010). A nonparametric algorithm for clustering microarray data. 45th *Scientific Meeting of the Italian Statistical Society*, Padova, Italy, June 16th – 18th, 2010.

De Bin, R., Risso, D. (2010). Clustering via nonparametric density estimation: an application to microarray data. (poster) *The* 29th *Leeds Annual Statistical Research Workshop*, Leeds, UK, July 6th – 8th, 2010.

Teaching experience

April 2010 – May 2010
Statistics II
Bachelor Degree in Economics
Teaching assistant, 15 hours
University of Ca' Foscari, Venice
Instructor: Prof. Francesca Parpinel

February 2010 – March 2010
Statistics I
Bachelor Degree in Economics
Teaching assistant, 15 hours
University of Ca' Foscari, Venice
Instructor: Prof. Andrea Pastore

References

Prof. Alessandra Salvan
Department of Statistical Sciences, University of Padova
via Cesare Battisti 241-243, 35121 Padova (PD), Italy
Phone: (+39) 049 827 4166
e-mail: salvan@stat.unipd.it

Prof. Nicola Sartori
Department of Statistical Sciences, University of Padova
via Cesare Battisti 241-243, 35121 Padova (PD), Italy
Phone: (+39) 049 827 4127
e-mail: sartori@stat.unipd.it

Prof. Thomas Severini
Department of Statistics, Northwestern University
2006 Sheridan Road, 60208 Evanston (IL), USA
Phone: (+1) 847 491 3974
e-mail: severini@northwestern.edu

Prof. Giovanni Andreatta
Department of Pure and Applied Mathematics, University of Padova
via Trieste 63, 35121 Padova (PD), Italy
Phone: (+39) 049 827 1483
e-mail: giovanni@math.unipd.it