Biometrika (2003), **90**, 3, *pp*. 533–549
© 2003 Biometrika Trust
Printed in Great Britain

Modified profile likelihoods in models with stratum nuisance parameters

By N. SARTORI

Department of Statistics, University of Padova, via C. Battisti 241–243, 35121 Padova, Italy
sartori@stat.unipd.it

Summary

It is well known, at least through many examples, that when there are many nuisance parameters modified profile likelihoods often perform much better than the profile likelihood. Ordinary asymptotics almost totally fail to deal with this issue. For this reason, we study asymptotic properties of the profile and modified profile likelihoods in models for stratified data in a two-index asymptotics setting. This means that both the sample size of the strata, *m*, and the dimension of the nuisance parameter, *q*, may increase to infinity. It is shown that in this asymptotic setting modified profile likelihoods give improvements, with respect to the profile likelihood, in terms of consistency of estimators and of asymptotic distributional properties. In particular, the modified profile likelihood based statistics have the usual asymptotic distribution, provided that $1/m = o(q^{-1/3})$, while the analogous condition for the profile likelihood is $1/m = o(q^{-1})$.

Some key words: Conditional likelihood; Incidental nuisance parameter; Modified profile likelihood; Profile likelihood; Profile score bias; Two-index asymptotics.

1. Introduction

Consider a model with a parameter θ written in the form $\theta = (\psi, \lambda)$, where ψ is a parameter of interest and λ is a nuisance parameter. Standard likelihood inference for ψ is usually based on the profile likelihood, which is the likelihood with the nuisance parameter replaced by its maximum likelihood estimate when the values of ψ are treated as fixed. The profile likelihood has some, but not all, of the properties of a proper likelihood. In particular, the expected value of the profile score is not zero. This defect may become a problem when the dimension of the nuisance parameter is large, relative to the sample size.

In problems with a substantial number of nuisance parameters, widely studied modified profile loglikelihoods of the form $l_{\mathrm{M}}(\psi) = l_{\mathrm{P}}(\psi) + M(\psi)$, where $l_{\mathrm{P}}(\psi)$ is the profile loglikelihood, may greatly improve inferences. Some examples of the function $M(\psi)$ are those of Barndorff-Nielsen (1983, 1994, 1995), Cox & Reid (1987), McCullagh & Tibshirani (1990) and Severini (1998). Often, this function is very simple, involving only the information matrix for nuisance parameters, which is ordinarily available from the computation of the profile loglikelihood. Typically, the modification $M(\psi)$ has the effect of reducing the profile score bias (DiCiccio et al., 1996).

Many have shown through examples that modified profile likelihoods can perform well when the profile likelihood fails. Ordinary asymptotics almost totally fail to deal with this issue since, there, likelihood-based inferences based on either are valid to first-order with no formal improvement for modified profile likelihoods. A way of dealing with this defect is to consider asymptotics where the number of nuisance parameters increases along with the sample size. What we show is that this increase can be much more rapid for first-order use of modified profile likelihoods than for first-order use of the profile likelihood.

In the following, we will refer to models with independent stratified observations of the form

$$Y_{ij} \sim p_{ij}(y_{ij}; \psi, \lambda_i),$$

where $i = 1, \ldots, q$ and $j = 1, \ldots, m$. The sample size is $n = mq$ and the nuisance parameter $\lambda = (\lambda_1, \ldots, \lambda_q)$ has dimension q. To study situations in which the number of nuisance parameters is rather large, it is useful to consider a two-index asymptotic setting, in which both the number of strata, q, and the stratum sample size, m, are allowed to increase to infinity. The two-index asymptotics setting will be called $(m \times q)$-asymptotics, as defined in Barndorff-Nielsen (1996) and also suggested in Davison (1992). The study of models where the dimension of the parameter may increase with the sample size has also been previously analysed in Portnoy (1988), Pierce & Peters (1992, § 3) and Barndorff-Nielsen & Cox (1994, § 8.5).

We note that, if q is fixed, we have a nuisance parameter of fixed dimension and the standard asymptotic theory applies. On the contrary, if m is fixed, we have $q = O(n)$ and maximum likelihood estimators are likely not to be consistent, as has been well known since Neyman & Scott (1948). In some cases, it is possible to solve this problem by using some inferential separation in the likelihood, as with conditional or marginal likelihoods. One notable example is the conditional likelihood for components of the canonical parameter in a full exponential family (Andersen, 1970, 1971; Davison, 1988). However, while conditional or marginal likelihoods are available only when the model has a particular structure, profile and modified profile likelihoods are general tools for inference. Moreover, modified profile likelihoods often provide accurate approximations to conditional or marginal likelihoods, when these exist.

In this paper, we give sufficient conditions for $l_{\mathrm{P}}(\psi)$ and $l_{\mathrm{M}}(\psi)$ to lead to ordinary chi-squared approximations for likelihood ratio tests in $(m \times q)$-asymptotics. When ψ is scalar, we consider the asymptotic normality of the signed squared root of the profile and modified profile likelihood ratio tests, denoted by $r_{\mathrm{P}}(\psi)$ and $r_{\mathrm{M}}(\psi)$ respectively. The condition for $l_{\mathrm{P}}(\psi)$ is already known, while the weaker condition for $l_{\mathrm{M}}(\psi)$ is the main result here. The argument is given in terms of models involving stratum nuisance parameters but, with little doubt, more general results of a similar nature could be obtained.

The main result is that, while methods based on the profile likelihood may fail unless $1/m = o(q^{-1})$, methods based on modified profile likelihoods still perform accurately, provided that $1/m = o(q^{-1/3})$. The result about the profile likelihood was already known in the context of regular exponential families (Portnoy, 1988), although expressed in a different form. Indeed, since $n = mq$, the above conditions can be written equivalently as $q = o(n^{1/2})$ and $q = o(n^{3/4})$, respectively.

The result about the profile likelihood is useful for identifying situations in which standard methods may fail, and it gives a motivation for the use, in practice, of marginal and conditional or modified profile likelihoods. The sufficient condition $1/m = o(q^{-1/3})$ indicates when inferences based on modified profile likelihoods are reliable.

Section 2 gives some notation and preliminaries. Section 3 shows the different $(m \times q)$-asymptotic properties of the profile and modified profile score statistics and § 4 deals with related likelihood quantities. Section 5 contains examples and simulations. Finally, a brief discussion is given in § 6.

2. Notation and preliminaries

We will consider models with independent stratified observations of the form

$$Y_{ij} \sim p_{ij}(y_{ij};\, \psi, \lambda_i),$$

where $i=1,\ldots,q$ and $j=1,\ldots,m_i$. The sample size is $n=\sum_{i=1}^q m_i$. We allow the dependence of the model function on i and j to include situations with explanatory variables x_{ij}. In such cases, we can write $p_{ij}(y_{ij};\, \psi, \lambda_i)=p(y_{ij};\, \psi, \lambda_i, x_{ij})$. However, the dependence of p on i and j will be dropped from notation in the following. We assume for simplicity that ψ and λ_i are scalars, although this assumption is not crucial for the results in the paper. Moreover, we also assume that $m_i=m$, but we could alternatively assume that each m_i can be written in the form $m_i=K_i m$, with $A \leqslant K_i \leqslant B$ and where A and B are positive finite numbers.

The loglikelihood can be written as

$$l(\theta)=\sum_{i=1}^{q} l^i(\psi, \lambda_i), \tag{1}$$

where

$$l^i(\psi, \lambda_i)=\sum_{j=1}^{m} \log p(y_{ij};\, \psi, \lambda_i)$$

is the loglikelihood function related to the ith stratum. We will assume usual regularity conditions for the loglikelihoods $l^i(\psi, \lambda_i)$; see for instance Severini (2000, § 3.4). Further regularity conditions will be stated below.

The maximum likelihood estimator of $\theta=(\psi, \lambda)$ will be denoted by $\hat\theta=(\hat\psi, \hat\lambda)$. The score function will be denoted by $U(\theta)=\partial l(\theta)/\partial\theta$, while the observed and expected information will be indicated by $j(\theta)$ and $i(\theta)$, respectively. For single components we will use subscripts. Moreover, in some cases, arguments will be dropped from the notation to avoid messy expressions; for example, we will use $U_\psi=U_\psi(\theta)=\partial l(\theta)/\partial\psi$, $j_{\psi\lambda_i}=j_{\psi\lambda_i}(\theta)=-\partial U_\psi(\theta)/\partial\lambda_i$ and $i_{\psi\lambda_i}=i_{\psi\lambda_i}(\theta)=E_\theta(j_{\psi\lambda_i})$.

As can be seen from (1), the loglikelihood is separable with respect to nuisance parameters, in that it is the sum of q terms, each of which depends on only one nuisance parameter. This is because the nuisance parameter λ_i is related just to the ith stratum and the strata are independent. This implies that, for each nuisance parameter, we can consider the standard asymptotics in powers of m. For instance, we have $U_{\lambda_i}(\theta)=U^i_{\lambda_i}(\psi, \lambda_i)=O_p(m^{1/2})$, $j_{\lambda_i\lambda_i}(\theta)=j^i_{\lambda_i\lambda_i}(\psi, \lambda_i)=O_p(m)$ and $j_{\lambda_i\lambda_k}(\theta)=0$, when $i \neq k$. Moreover, the separability of nuisance parameters also implies that the constrained maximum likelihood estimate of λ for fixed ψ, $\hat\lambda_\psi$, is the solution of the q independent likelihood equations for the strata. As a consequence, the profile loglikelihood for ψ may be written as the sum of the q profile loglikelihoods for the strata:

$$l_{\mathrm{P}}(\psi)=l(\psi, \hat\lambda_\psi)=\sum_{i=1}^{q} l^i(\psi, \hat\lambda_{i\psi})=\sum_{i=1}^{q} l^i_P(\psi). \tag{2}$$

In the following, U_{P} will denote the profile score function. After standard expansions, such as those in the Appendix of McCullagh & Tibshirani (1990), we have the following expansion for the profile score in the ith stratum:

$$U_P^i = U_{\psi|\lambda_i} + B_P^i + R_P^i, \tag{3}$$

where $U_{\psi|\lambda_i} = U_\psi^i - i_{\psi\lambda_i} i_{\lambda_i\lambda_i}^{-1} U_{\lambda_i}$, B_P^i is a term of order $O_p(1)$ with expected value $-\rho_\psi^i$, of order $o(1)$, and R_P^i is the remainder term of order $O_p(m^{-1/2})$ and with expected value of order $O(m^{-1})$; see also DiCiccio et al. (1996). Hence, we have that the bias of the profile score in the ith stratum is $E_\theta(U_P^i) = -\rho_\psi^i + O(m^{-1})$. As noted also by McCullagh & Tibshirani (1990), the problem with the profile likelihood for stratified data is that the bias of the profile score accumulates across strata. In fact, from (2), it is straightforward to see that the profile score bias has leading term $-\sum_{i=1}^q \rho_\psi^i = O(q)$.

In the following, we will consider modified profile loglikelihoods of the form

$$l_M(\psi) = l_P(\psi) + M(\psi),$$

where the modification $M(\psi)$ is a suitably smooth function having derivatives of order $O_P(1)$. Some examples of the modification $M(\psi)$ are given in McCullagh & Tibshirani (1990), Cox & Reid (1987) and Barndorff-Nielsen (1994, 1995). Throughout the paper, we will mainly refer to the modified profile likelihood of Barndorff-Nielsen (1983), which is a highly accurate approximation to a conditional or marginal likelihood, when either exists. Moreover, it is also invariant with respect to interest-respecting reparameterisations. It has a modification of the form

$$M(\psi) = \tfrac{1}{2}\log|j_{\lambda\lambda}(\hat\theta_\psi)| - \log|l_{\lambda;\hat\lambda}(\hat\theta_\psi)|$$

with $l_{\lambda;\hat\lambda}(\theta) = \partial^2 l(\theta; \hat\theta, a)/(\partial\lambda\, \partial\hat\lambda^T)$. The computation of this term requires a sample space derivative. This means that we need to write the data y in the form $(\hat\theta, a)$, where a is an ancillary statistic, either exactly or approximately. This is straightforward in full exponential families and in transformation models. For a general model, we can use second-order approximations for $M(\psi)$ which are based on approximations to required sample space derivatives; see Severini (2000, § 9.5) for a recent review.

Example 1: *Full exponential family*. Consider a full exponential family with a component of the canonical parameter as the parameter of interest. With stratified data, the loglikelihood can be written in the form

$$l(\theta) = \sum_{i=1}^q l^i(\psi, \lambda_i) = \sum_{i=1}^q \{\psi u_i + \lambda_i v_i - mK(\psi, \lambda_i)\},$$

where the sufficient statistic in the ith stratum has components $u_i = \sum_j u_{ij}$ and $v_i = \sum_j v_{ij}$ and $K(.)$ is the cumulant function. The overall sufficient statistic has components $u = \sum_{i=1}^q u_i$ and $v = (v_1, \ldots, v_q)$.

The modified profile loglikelihood is an approximation for the conditional loglikelihood for ψ, and it has

$$M(\psi) = \frac{1}{2}\log|K_{\lambda\lambda}(\psi, \hat\lambda_\psi)| = \sum_{i=1}^q M^i(\psi) = \sum_{i=1}^q \frac{1}{2}\log K_{\lambda_i\lambda_i}(\psi, \hat\lambda_{i\psi}),$$

where $K_{\lambda\lambda} = \partial^2 K/(\partial\lambda\, \partial\lambda^T)$, and is exactly the sum of the modified profile loglikelihoods of each single stratum. This may not be the case in general models. We note that, in this case, $M(\psi)$ involves only the information matrix for λ. This happens also for location-scale regression models, as noted in the reply to the discussion in Pierce & Peters (1992).

3. Score statistics

In this section we consider the $(m \times q)$-asymptotic distribution of the profile and modified profile score statistics

$$W^u_{\mathrm{P}}(\psi) = j_{\mathrm{P}}(\psi)^{-1} U_{\mathrm{P}}(\psi)^2, \quad W^u_{\mathrm{M}}(\psi) = j_{\mathrm{M}}(\psi)^{-1} U_{\mathrm{M}}(\psi)^2,$$

where j_{P} and j_{M} are the profile and modified profile observed information, and U_{M} is the modified profile score function.

We note that, in stratified models, the separability of nuisance parameters implies that the modification $M(\psi)$ satisfies

$$M(\psi) = \sum_{i=1}^{q} M^i(\psi), \tag{4}$$

where $M^i(\psi)$ is such that

$$E_\theta\left(\frac{\partial M^i}{\partial \psi}\right) = -E_\theta(U^i_{\mathrm{P}}) + O(m^{-1}) = \rho^i_\psi + O(m^{-1}). \tag{5}$$

It is well known, see for example DiCiccio et al. (1996), that, for any of the proposed modified profile likelihoods, the ρ^i_ψ in (5) is the same as discussed following (3), and hence for each stratum the modifications eliminate all but $O(m^{-1})$ of the profile score bias. Clearly this also holds in asymptotics involving a fixed number of strata, and in the following we investigate the implications for asymptotics where the number of strata increases.

Consider the following expansions for the profile and modified profile score functions:

$$U_{\mathrm{P}} = U_{\psi|\lambda} + B_{\mathrm{P}} + R_{\mathrm{P}}, \tag{6}$$

$$\begin{aligned} U_{\mathrm{M}} &= U_{\psi|\lambda} + B_{\mathrm{M}} + R_{\mathrm{P}} \\ &= U_{\psi|\lambda} + R_{\mathrm{M}}. \end{aligned} \tag{7}$$

Expansion (6) follows directly from (2) and (3) if we write

$$U_{\psi|\lambda} = U_\psi - i_{\psi\lambda} i_{\lambda\lambda}^{-1} U_\lambda = \sum_{i=1}^{q} U_{\psi|\lambda_i},$$

$B_{\mathrm{P}} = \sum_{i=1}^q B^i$ and $R_{\mathrm{P}} = \sum_{i=1}^q R^i_{\mathrm{P}}$, while expansion (7) follows from $U_{\mathrm{M}} = U_{\mathrm{P}} + \partial M/\partial\psi$ and (6). In particular, $B_{\mathrm{M}} = B_{\mathrm{P}} + \partial M/\partial\psi$ and $R_{\mathrm{M}} = B_{\mathrm{M}} + R_{\mathrm{P}}$ are quantities of the same order as R_{P}, as explained below.

The quantity $U_{\psi|\lambda}$ is essentially the score function for ψ when the nuisance parameters are considered as known. Note that its variance is $i_{\psi\psi|\lambda} = i_{\psi\psi} - i_{\psi\lambda} i_{\lambda\lambda}^{-1} i_{\lambda\psi}$. This quantity is also called the partial information for ψ (Severini, 2000, § 3.6.3) and, as with $U_{\psi|\lambda}$, is additive among strata. Indeed, independence between strata implies that

$$i_{\psi\psi|\lambda} = \sum_{i=1}^{q} V_\theta(U_{\psi|\lambda_i}) = \sum_{i=1}^{q} i_{\psi\psi|\lambda_i}.$$

In the following, we assume that the sequence of quantities $\bar{\imath}_{\psi\psi|\lambda} = i_{\psi\psi|\lambda}/n = i_{\psi\psi|\lambda}/(mq)$ converges to a strictly positive number, as q and m diverge. This reasonable assumption allows us to write $i_{\psi\psi|\lambda} = n\bar{\imath}_{\psi\psi|\lambda}$ and it guarantees an asymptotic lower bound for the partial information. Further, we assume the following: (i) the ordinary conditions that arise in the asymptotic setting with independent and identically distributed random variables hold;

(ii) the modification $M(\psi)$ satisfies properties (4) and (5); (iii) all relevant O_p bounds are uniform over strata so that, for example, if some $A_i = O(1)$ then $\sum_{i=1}^{q} A_i = O(q)$. In particular, in the presence of covariables, (ii) and (iii) are basically conditions on the sequence of covariables.

A central point here is that the quantity $Z = i_{\psi\psi|\lambda}^{-1/2} U_{\psi|\lambda}$ is asymptotically standard normal as n diverges, regardless of the nature of the sequence $\{q, m\}$. In fact, $U_{\psi|\lambda}$ is a sum of n independent quantities with zero mean and with the sum of the variances equal to $i_{\psi\psi|\lambda}$. Hence, as n goes to infinity and under standard regularity conditions, the asymptotic normality follows from the central limit theorem for independent variables. We note, however, that Z cannot be used for inference, since it depends on λ.

The central limit theorem for independent variables is also used to evaluate the order of the residual terms in (6) and (7). In particular, as shown in the Appendix, B_{P} is of order $O_p(q)$, as a consequence of the score bias, and $R_{\mathrm{P}} = O_{\mathrm{P}}[\max\{q/m, (q/m)^{1/2}\}]$. The latter expression is rather unconventional but it is required by the unconventional $(m \times q)$-asymptotics, with the two terms corresponding respectively to whether it is the mean or the standard deviation of R_{P} that dominates. For what concerns U_{M}, the crucial point is that B_{M} is of the same order as R_{P}. The reason for this is that the adjustment $\partial M/\partial\psi$ removes a major part of the profile score bias, due to B_{P}, leaving only a part that is of the same order as R_{P}. This is because of properties (4) and (5), together with expansion (6), as shown in the Appendix.

Consider now the standardised score functions, using temporarily the partial information in place of the observed information. From (6) and (7) we obtain

$$i_{\psi\psi|\lambda}^{-1/2} U_{\mathrm{P}} = Z + i_{\psi\psi|\lambda}^{-1/2} B_{\mathrm{P}} + i_{\psi\psi|\lambda}^{-1/2} R_{\mathrm{P}}, \quad i_{\psi\psi|\lambda}^{-1/2} U_{\mathrm{M}} = Z + i_{\psi\psi|\lambda}^{-1/2} R_{\mathrm{M}},$$

where, in the latter case, the term $i_{\psi\psi|\lambda}^{-1/2} B_{\mathrm{P}} = O_p\{(q/m)^{1/2}\}$ has been removed by the bias correction of the modified profile likelihood. On the other hand, both residual terms, $i_{\psi\psi|\lambda}^{-1/2} R_{\mathrm{P}}$ and $i_{\psi\psi|\lambda}^{-1/2} R_{\mathrm{M}}$, are of order $O_p[\max\{(q/m^3)^{1/2}, m^{-1}\}] = o_p\{(q/m)^{1/2}\}$.

For the asymptotic normality of the standardised score statistics we need Z to be the leading term in the expansions above, which means that the remaining terms need to be $o_p(1)$. In the first case, we have

$$i_{\psi\psi|\lambda}^{-1/2} U_{\mathrm{P}} = Z + O_p\{(q/m)^{1/2}\} = Z + O_p(qn^{-1/2}). \tag{8}$$

This implies that a sufficient condition for $i_{\psi\psi|\lambda}^{-1/2} U_{\mathrm{P}}$ to be asymptotically standard normal is that $q/m = o(1)$. Hence, we require the sample size in each stratum to increase at a faster rate than the number of strata. The condition can be written in the equivalent forms $1/m = o(q^{-1})$ and $q = o(n^{1/2})$. On the other hand, for a modified profile likelihood,

$$i_{\psi\psi|\lambda}^{-1/2} U_{\mathrm{M}} = Z + O_p[\max\{(q/m^3)^{1/2}, m^{-1}\}].$$

This implies that, in this case, the condition becomes $q/m^3 = o(1)$. This may be stated in different, but equivalent, forms, such as $1/m = o(q^{-1/3})$ or $q = o(n^{3/4})$. In particular, the condition $1/m = o(q^{-1/3})$ shows that, if we let q increase, m has to increase as well, and it has to increase faster than $q^{1/3}$. From another point of view, the equivalent condition $q = o(n^{3/4})$ says that q can increase with n, but it has to increase more slowly than $n^{3/4}$. In any case, the point is that the condition required for the modified profile likelihood is weaker than that required for the profile likelihood. Thus, in situations in which m increases faster than $q^{1/3}$, but not faster than q, $i_{\psi\psi|\lambda}^{-1/2} U_{\mathrm{M}}$ has the usual asymptotic distribution while this cannot be guaranteed for $i_{\psi\psi|\lambda}^{-1/2} U_{\mathrm{P}}$.

We note that the results are the same whether we use $i_{\psi\psi|\lambda}$ or the observed information.

Indeed, in formula (A3) in the Appendix, it is shown that $j_{\mathrm{P}} = i_{\psi\psi|\lambda}\{1 + o_p(1)\}$, so that

$$j_{\mathrm{P}}^{-1/2} U_{\mathrm{P}} = i_{\psi\psi|\lambda}^{-1/2} U_{\mathrm{P}}\{1 + o_p(1)\},$$

where the relative error is of order $O_p(n^{-1/2})$ when $1/m = o(q^{-1})$ and $O_p(m^{-1})$ otherwise. Hence, also $j_{\mathrm{P}}^{-1/2} U_{\mathrm{P}}$ is asymptotically standard normal, provided that $1/m = o(q^{-1})$. This also implies that $W_{\mathrm{P}}^{u} = Z^2 + o_p(1)$, giving the usual χ_1^2 asymptotic distribution. Analogously, we have that

$$j_{\mathrm{M}}^{-1/2} U_{\mathrm{M}} = i_{\psi\psi|\lambda}^{-1/2} U_{\mathrm{M}}\{1 + o_p(1)\}.$$

Hence, $j_{\mathrm{M}}^{-1/2} U_{\mathrm{M}}$ is asymptotically standard normal, and W_{M}^{u} asymptotically χ_1^2, provided that $1/m = o(q^{-1/3})$.

As a final remark, note that, when $1/m = o(q^{-1})$, we have that

$$j_{\mathrm{M}}^{-1/2} U_{\mathrm{M}} = Z + O_p(n^{-1/2})$$

as opposed to (8). Hence, even in situations when both profile and modified profile score statistics have the usual asymptotic distribution, the latter has a smaller upper bound for the error term.

4. Other likelihood quantities

Here, we will see how the results of § 3 bear on consistency of estimators and the asymptotic distribution of other likelihood-based statistics, such as the likelihood ratio and Wald statistics. It is shown that the three versions of likelihood-based statistics are asymptotically equivalent, as in the standard asymptotic setting. For the estimators, it is shown that modified profile likelihoods give improvements in terms of consistency over the profile likelihood.

In general, denoting by $\hat{\psi}_{\mathrm{M}}$ the maximiser of l_{M}, we have that $\hat{\psi}$ and $\hat{\psi}_{\mathrm{M}}$ will be consistent, no matter what is the nature of the sequence $\{m, q\}$, since both m and q go to infinity. However, the rates of convergence to the true parameter value depend on the relative rates of q and m.

Let us consider the profile case. An expansion for the profile likelihood equation around ψ gives

$$0 = U_{\mathrm{P}}(\hat{\psi}) = U_{\mathrm{P}}(\psi) - j_{\mathrm{P}}(\psi)(\hat{\psi} - \psi) + O_p(n)O_p\{(\hat{\psi} - \psi)^2\},$$

where we assume that third and subsequent derivatives of the profile loglikelihood are of order $O_p(n)$. This implies that

$$(\hat{\psi} - \psi) = j_{\mathrm{P}}(\psi)^{-1} U_{\mathrm{P}}(\psi) + O_p\{(\hat{\psi} - \psi)^2\}. \qquad (9)$$

From (6) and (A3), it is straightforward to see that

$$\hat{\psi} = \psi + O_p(n^{-1/2})$$

if $1/m = o(q^{-1})$, while $\hat{\psi} = \psi + O_p(m^{-1})$ otherwise. On the other hand, using analogous reasoning for the modified profile likelihood, we have that $\hat{\psi}_{\mathrm{M}} = \psi + O_p(n^{-1/2})$ when $1/m = o(q^{-1/3})$, and $\hat{\psi}_{\mathrm{M}} = \psi + O_p(m^{-2})$ otherwise. Hence, when q increases faster than m, $\hat{\psi}$ might converge to ψ at a slower rate than $\hat{\psi}_{\mathrm{M}}$. In some sense, this formalises the comment in Barndorff-Nielsen & Cox (1994, p. 285) that, in $(m \times q)$-asymptotics, $\hat{\psi}_{\mathrm{M}}$ 'will generally be more nearly consistent than $\hat{\psi}$'.

Consider now the likelihood ratio and Wald statistics based on profile and modified profile likelihoods. In particular, the likelihood ratio statistics are

$$W_{\mathrm{P}}(\psi) = 2\{l_{\mathrm{P}}(\hat{\psi}) - l_{\mathrm{P}}(\psi)\}, \quad W_{\mathrm{M}}(\psi) = 2\{l_{\mathrm{M}}(\hat{\psi}_{\mathrm{M}}) - l_{\mathrm{M}}(\psi)\},$$

and the Wald statistics are

$$W^e_{\mathrm{P}}(\psi) = j_{\mathrm{P}}(\psi)(\hat{\psi} - \psi)^2, \quad W^e_{\mathrm{M}}(\psi) = j_{\mathrm{M}}(\psi)(\hat{\psi}_{\mathrm{M}} - \psi)^2.$$

In the standard asymptotic setting, score, Wald and likelihood ratio statistics are first-order asymptotically equivalent. It is possible to see that this is still true in $(m \times q)$-asymptotics. Indeed, if we use (9) and the above results about $(\hat{\psi} - \psi)$, it follows that

$$W^e_{\mathrm{P}} = W^u_{\mathrm{P}}\{1 + O_p(n^{-1/2})\},$$

provided that $1/m = o(q^{-1})$. If this condition does not hold, the above relationship becomes $W^e_{\mathrm{P}} = W^u_{\mathrm{P}}\{1 + O_p(m^{-1})\}$. Similarly, expanding $l_{\mathrm{P}}(\psi)$ around $\hat{\psi}$, we obtain

$$W_{\mathrm{P}} = W^e_{\mathrm{P}}\{1 + O_p(n^{-1/2})\},$$

when $1/m = o(q^{-1})$, and $W_{\mathrm{P}} = W^e_{\mathrm{P}}\{1 + O_p(m^{-1})\}$ otherwise. This implies that W_{P} and W^e_{P} asymptotically have χ^2_1 distributions, as does W^u_{P}, provided that $1/m = o(q^{-1})$.

For the modified profile likelihood, following the same steps we obtain

$$W^e_{\mathrm{M}} = W^u_{\mathrm{M}}\{1 + O_p(n^{-1/2})\}, \quad W_{\mathrm{M}} = W^e_{\mathrm{M}}\{1 + O_p(n^{-1/2})\},$$

when $1/m = o(q^{-1/3})$, while the relative error is of order $O_p(m^{-1})$ otherwise. This means that W_{M} and W^e_{M} have χ^2_1 asymptotic distributions, provided that $1/m = o(q^{-1/3})$, as for W^u_{M}.

What has been shown is that the three likelihood-based statistics are asymptotically equivalent to $o_p(1)$. This means that, when one of them has the usual asymptotic distribution, the other two are equivalent to it with a relative error of order $O_p(n^{-1/2})$, as in standard asymptotics. Also the reverse is true, in that when one fails so do the other two. The conditions for the usual asymptotic results are those found in the previous section.

As a final remark, note that, in the formulae for the score and Wald statistics, we can alternatively use the observed information evaluated at the estimator of ψ, obtaining the same asymptotic results. This is easily shown by expanding the observed information around the maximum likelihood estimator.

5. Examples

Example 1: *Full exponential family* (*cont.*). In full exponential families, a conditional likelihood for the canonical parameter ψ is available, at least in principle, namely the likelihood associated with the conditional distribution of U given V. Hence, profile and modified profile likelihoods can be also compared with this conditional likelihood. In particular, the conditional loglikelihood is given by

$$l_{\mathrm{C}}(\psi) = l(\theta) - l(\theta; v),$$

where $l(\theta; v)$ is the loglikelihood based on the marginal distribution of V. In the standard asymptotic setting, the conditional likelihood can be approximated using a saddlepoint approximation for the marginal distribution of V (Severini, 2000, § 8.2.4). This gives the modified profile likelihood with an error term of order $O(n^{-1})$. In $(m \times q)$-asymptotics, we can apply similar reasoning. First, note that the conditional loglikelihood for ψ can be written in the form

$$l_{\mathrm{C}}(\psi)=\sum_{i=1}^{q} l_{\mathrm{C}}^{i}(\psi), \tag{10}$$

where $l_{\mathrm{C}}^{i}(\psi)$ is the conditional loglikelihood obtained from the distribution of U_i given V_i. The additivity property holds for the profile and the modified profile loglikelihood as well, as already seen in § 2. Moreover, in each stratum, it is well known that we have

$$\begin{aligned} l_{\mathrm{C}}^{i}(\psi)&=l_{\mathrm{P}}^{i}(\psi)+O_p(1),\\ l_{\mathrm{C}}^{i}(\psi)&=l_{\mathrm{P}}^{i}(\psi)+M^{i}(\psi)+O_p(m^{-1}). \end{aligned} \tag{11}$$

Note that $M^{i}(\psi)$ and its derivatives are of order $O_p(1)$, and derivatives of the residual term in (11) are also of order $O_p(m^{-1})$. This implies that $l_{\mathrm{C}}(\psi)=l_{\mathrm{P}}(\psi)+O_p(q)$ and $l_{\mathrm{C}}(\psi)=l_{\mathrm{M}}(\psi)+O_p(q/m)$. However, the relationship between $l_{\mathrm{P}}(\psi)$, $l_{\mathrm{M}}(\psi)$ and $l_{\mathrm{C}}(\psi)$ is better explained when we consider the relative versions of these loglikelihoods. We will denote the relative loglikelihoods by $\bar{l}_{\mathrm{P}}(\psi)$, $\bar{l}_{\mathrm{M}}(\psi)$ and $\bar{l}_{\mathrm{C}}(\psi)$. After some standard expansions it is possible to prove that

$$\begin{aligned} \bar{l}_{\mathrm{P}}(\psi)&=\bar{l}_{\mathrm{C}}(\psi)+o_p(1) \quad (1/m=o(q^{-1})),\\ \bar{l}_{\mathrm{M}}(\psi)&=\bar{l}_{\mathrm{C}}(\psi)+o_p(1) \quad (1/m=o(q^{-1/3})). \end{aligned}$$

These results generalise those of Barndorff-Nielsen (1996) regarding the gamma distribution and are in agreement with those of § 4. In fact, the conditional likelihood ratio statistic, $W_{\mathrm{C}}(\psi)=-2\bar{l}_{\mathrm{C}}(\psi)$, asymptotically has a χ_1^2 distribution, as $q\rightarrow\infty$. This is true even in the case with m fixed (Andersen, 1971). The same result holds for $W_{\mathrm{P}}(\psi)$, if $1/m=o(q^{-1})$, and for $W_{\mathrm{M}}(\psi)$, if $1/m=o(q^{-1/3})$.

A special case is when the Y_{ij} are independent normal with mean λ_i and variance ψ. Indeed, the modified profile likelihood for ψ is exactly equal to the conditional likelihood, which is also a marginal likelihood; see for instance Example 9.6 in Severini (2000). The same happens when Y_{ij} are independent inverse Gaussian random variables; see Example 1 in Sartori et al. (1999). Examples 2 and 3 below are instances where the conditional likelihood exists, but is not equal to the modified profile likelihood.

Example 2: *Gamma samples with common shape parameter.* Suppose the Y_{ij} are independent gamma random variables with shape parameter ψ and scale parameter $1/\lambda_i$, as in Example 5.1 of Barndorff-Nielsen (1996). The sufficient statistic has components $u=\sum_i^q\sum_j^m \log y_{ij}$ and $v_i=\sum_j^m y_{ij}$ $(i=1,\ldots,q)$. If we write $s=u-m\sum_i^q \log v_i$, the loglikelihoods are

$$\begin{aligned} l_{\mathrm{C}}(\psi)&=\psi s+q\log\Gamma(m\psi)-mq\log\Gamma(\psi),\\ l_{\mathrm{P}}(\psi)&=\psi s+mq\psi\log m\psi-mq\psi-mq\log\Gamma(\psi),\\ l_{\mathrm{M}}(\psi)&=\psi s+q(m\psi-0{\cdot}5)\log m\psi-mq\psi-mq\log\Gamma(\psi), \end{aligned}$$

which have to be maximised numerically. Denote by r_{C}, r_{P} and r_{M} the signed square roots of the conditional, profile and modified profile likelihood ratio statistics. Table 1 reports the probabilities $\Phi\{r_{\mathrm{P}}(\psi)\}$ and $\Phi\{r_{\mathrm{M}}(\psi)\}$ for several values of m and q, where $\Phi(.)$ is the cumulative distribution function of the standard normal. For each combination of m and q, ψ and s are such that $\Phi\{r_{\mathrm{C}}(\psi)\}=0{\cdot}05$ and $\hat{\psi}_{\mathrm{C}}=1$. The numerical results confirm the theoretical ones. As an indication, although arbitrary, for each value of q, we put in bold face the cell corresponding to the smallest value of m that gives a probability within 0·01

Table 1: *Example* 2. *Inference about common shape parameter in q gamma samples of size m. Probabilities* $\Phi\{r_{\mathrm{P}}(\psi)\}$ *and* $\Phi\{r_{\mathrm{M}}(\psi)\}$ *with* ψ *such that* $\Phi\{r_{\mathrm{C}}(\psi)\}=0{\cdot}05$ *in samples with* $\hat{\psi}_{\mathrm{C}}=1$. *For each q, values in bold face correspond to the smallest m, for* r_{P} *and* r_{M}, *such that the probability is within* 0·01 *of* 0·05.

m		$q=4$	$q=8$	$q=16$	$q=64$	$q=128$
3	r_{P}	0·190	0·299	0·487	0·952	0·999
	r_{M}	**0·053**	**0·055**	**0·058**	0·070	0·080
4	r_{P}	0·159	0·239	0·383	0·866	0·988
	r_{M}	0·052	0·053	0·055	0·062	0·068
5	r_{P}	0·141	0·206	0·322	0·777	0·962
	r_{M}	0·052	0·052	0·054	**0·058**	0·062
6	r_{P}	0·129	0·184	0·281	0·698	0·924
	r_{M}	0·051	0·052	0·053	0·056	**0·059**
200	r_{P}	**0·060**	0·064	0·071	0·098	0·126
	r_{M}	0·050	0·050	0·050	0·050	0·050
400	r_{P}	0·056	**0·059**	0·064	0·081	0·098
	r_{M}	0·050	0·050	0·050	0·050	0·050
800	r_{P}	0·054	0·057	**0·059**	0·071	0·080
	r_{M}	0·050	0·050	0·050	0·050	0·050
3000	r_{P}	0·052	0·053	0·055	**0·059**	0·064
	r_{M}	0·050	0·050	0·050	0·050	0·050
6000	r_{P}	0·051	0·053	0·053	0·057	**0·060**
	r_{M}	0·050	0·050	0·050	0·050	0·050

of the target value. The modified profile likelihood gives practically the same results as the conditional, never requiring m to be larger than 6. In contrast, the profile likelihood needs very large values of m, even for moderately large values of q.

Example 3: *Inference about common odds ratio in* 2 × 2 *tables.* Consider q independent pairs of independent binomial variables (Y_{i1}, Y_{i2}), with $Y_{i1} \sim \mathrm{Bi}(1, p_{i1})$ and $Y_{i2} \sim \mathrm{Bi}(m, p_{i2})$. With the parameterisation $\lambda_i = \log\{p_{i2}/(1-p_{i2})\}$ and

$$\psi = \log\{p_{i1}/(1-p_{i1})\} - \log\{p_{i2}/(1-p_{i2})\},$$

the model is again a full-rank exponential family, as in Example 1, with components of the sufficient statistic $u = \sum_i^q y_{i1}$ and $v_i = y_{i1} + y_{i2}$ $(i = 1, \ldots, q)$. This model may arise in case-control studies, in which we have one case and m controls in each table, and where interest is in studying the influence of some risk factor. The conditional likelihood is a noncentral hypergeometric distribution; see for instance Example 6.1 in Davison (1988). Profile and modified profile likelihood are easily computed using standard software for generalised linear models. Pierce & Peters (1992) studied in detail higher-order methods in this setting. Here, we propose only a particular example rather than giving a detailed exploration. The aim is to show the influence of m and q on the accuracy of the approximation to the conditional likelihood given by the profile and modified profile likelihoods. Table 2 reports the probabilities $\Phi\{r_{\mathrm{P}}(\psi_l)\}$, $1-\Phi\{r_{\mathrm{P}}(\psi_u)\}$ and $\Phi\{r_{\mathrm{M}}(\psi_l)\}$, $1-\Phi\{r_{\mathrm{M}}(\psi_u)\}$, where ψ_l and ψ_u are such that $\Phi\{r_{\mathrm{C}}(\psi_u)\} = 1-\Phi\{r_{\mathrm{C}}(\psi_l)\} = 0{\cdot}05$. In order to have comparable settings, we considered only odd values of m and tables with $v_i = (m+1)/2$, corre-

Table 2: *Example* 3. *Inference about common odds ratio in pairs of binomial observations. One-sided non-coverage probabilities of the* 0·90 *conditional confidence interval. We consider q tables with one case and m controls with $v_i = (m+1)/2$ in each table. We fix u equal to the closest integer that gives $\hat{\psi}_C = 1$. For each q, values in bold face correspond to the smallest m, for r_P and r_M, such that the probabilities are both within* 0·01 *of* 0·05.

m		$q=4$	$q=20$	$q=50$	$q=300$	$q=600$
1	r_P	0·030, 0·229	0·003, 0·595	0·000, 0·863	0·000, 1·000	0·000, 1·000
	r_M	**0·044, 0·056**	0·021, 0·127	0·001, 0·214	0·000, 0·653	0·000. 0·889
3	r_P	0·043, 0·084	0·019, 0·163	0·009, 0·262	0·000, 0·735	0·000, 0·937
	r_M	0·049, 0·046	**0·043, 0·051**	0·039, 0·058	0·025, 0·086	0·019, 0·108
5	r_P	0·046, 0·068	0·028, 0·107	0·018, 0·153	0·002, 0·407	0·000, 0·623
	r_M	0·050, 0·048	0·047, 0·050	**0·045, 0·052**	0·039, 0·061	0·035, 0·067
7	r_P	0·047, 0·063	0·033, 0·088	0·024, 0·115	0·005, 0·266	0·002, 0·415
	r_M	0·050, 0·049	0·048, 0·050	0·047, 0·051	**0·044, 0·055**	**0·042, 0·058**
9	r_P	**0·048, 0·060**	0·036, 0·078	0·028, 0·098	0·009, 0·199	0·004, 0·301
	r_M	0·050, 0·049	0·049, 0·050	0·048, 0·050	0·046, 0·053	0·045, 0·054
23	r_P	0·049, 0·054	**0·044, 0·060**	0·040, 0·066	0·027, 0·092	0·020, 0·116
	r_M	0·050, 0·050	0·050, 0·050	0·050, 0·050	0·049, 0·050	0·049, 0·051
35	r_P	0·049, 0·052	0·046, 0·056	**0·043, 0·060**	0·033, 0·076	0·028, 0·089
	r_M	0·050, 0·050	0·050, 0·050	0·050, 0·050	0·050, 0·050	0·050, 0·050
79	r_P	0·050, 0·051	0·048, 0·053	0·047, 0·054	**0·042, 0·060**	0·039, 0·065
	r_M	0·050, 0·050	0·050, 0·050	0·050, 0·050	0·050, 0·050	0·050, 0·050
111	r_P	0·050, 0·051	0·049, 0·052	0·048, 0·053	0·044, 0·057	**0·042, 0·060**
	r_M	0·050, 0·050	0·050, 0·050	0·050, 0·050	0·050, 0·050	0·050, 0·050

sponding to equal numbers of successes and failures. Moreover, we fixed u by assuming the condition $\hat{\psi}_C = 1$.

The results show very accurate behaviour of $r_M(\psi)$, even for large values of q and moderate values of m. As noted also in Pierce & Peters (1992), the accuracy of the approximations may depend also on the observed value of u. Indeed, u increases in steps of one from 0 to q. In the same setting as that of Table 2, the behaviour of the three likelihoods as functions of u is symmetric around $q/2$. Figure 1 shows conditional, profile and modified profile relative loglikelihoods in the case of $q = 100$, $m = 5$ and with values of u equal to 5, 30 and 45. The accuracy of the modified profile likelihood tends to be slightly worse when u approaches its boundary, while it is otherwise an accurate approximation of the conditional likelihood. On the contrary, the profile likelihood gives reasonable results only when u is close to its midrange.

Example 4: *Matched gamma pairs.* Let Y_{ij1} and Y_{ij2} be independent exponential random variables with means ψ/λ_i and $\psi\lambda_i$, respectively. This is equivalent to considering $Y_{i1} = \sum_{j=1}^{m} Y_{ij1}$ and $Y_{i2} = \sum_{j=1}^{m} Y_{ij2}$ as matched gamma pairs with shape m and scales λ_i/ψ and $1/(\psi\lambda_i)$, respectively. In this case, there is no exact conditional or marginal likelihood for ψ. Hence, we compare the profile likelihood and a modification of it, through a simulation study. In this problem, the modified profile likelihood coincides with the approximate conditional likelihood of Cox & Reid (1987), since $\hat{\lambda}_\psi = \hat{\lambda}$. The profile and

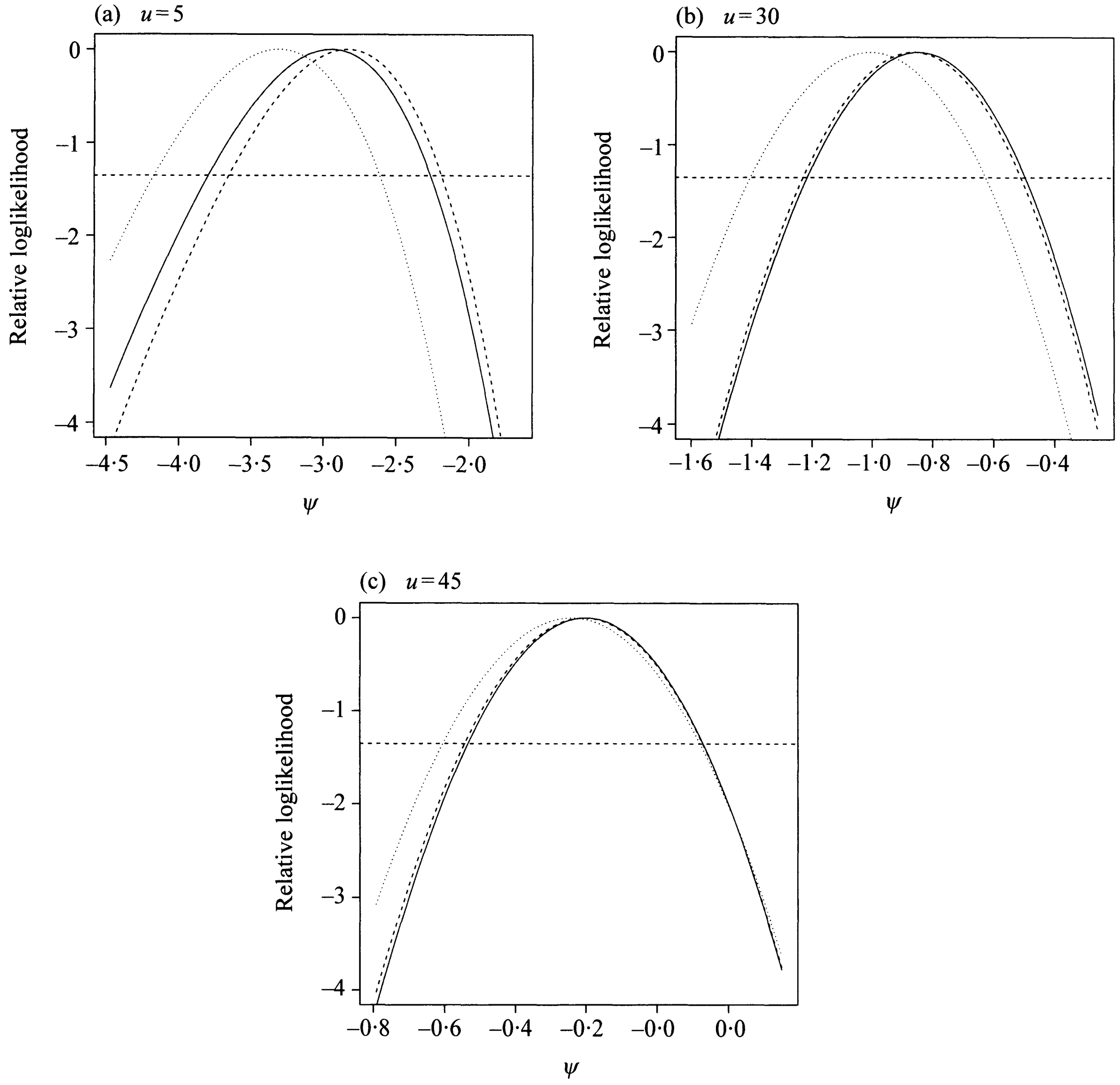

Fig. 1: Example 3. Inference about the common odds ratio in pairs of binomial observations. Relative loglikelihoods for ψ: profile (dotted), modified profile (dashed) and conditional (solid). The horizontal dashed line gives the 0·90 confidence interval. We consider $q=100$, $m=5$, $v_i=3$ $(i=1,\ldots,q)$, and three values of u.

modified profile loglikelihoods are

$$l_{\mathrm{P}}(\psi)=-2mq\log\psi-\frac{2}{\psi}\sum_{i=1}^{q}(y_{i1}y_{i2}),\quad l_{\mathrm{M}}(\psi)=l_{\mathrm{P}}(\psi)+\frac{1}{2}q\log\psi,$$

and the maximum likelihood estimators are

$$\hat{\psi}=(mq)^{-1}\sum_{i}(y_{i1}y_{i2})^{1/2},\quad \hat{\psi}_{\mathrm{M}}=\frac{4m}{4m-1}\hat{\psi}.$$

Cox & Reid (1992) show that $\hat{\psi}_{\mathrm{M}}$ is less biased than $\hat{\psi}$. Here, we compare the empirical distributions of $r_{\mathrm{P}}(\psi)$ and $r_{\mathrm{M}}(\psi)$ with the standard normal distribution, in simulations with various values of m and q. In this case, the actual stratum sample size is $2m$. As an example,

Table 3: *Example* 4. *Inference about common product of means in q pairs of gamma observations with shape m. Coverage probabilities of the* 0·05 *quantile of the standard normal distribution for* $r_P(\psi)$ *and* $r_M(\psi)$ *in simulations with* 100 000 *replications each, and with various values of m and q. The parameter of interest is* $\psi = 1$. *For each q, values in bold face correspond to the smallest m, for* r_P *and* r_M, *such that the probabilities are both within* 0·01 *of* 0·05.

m		$q=4$	$q=8$	$q=16$	$q=64$	$q=128$
1	r_P	0·210	0·278	0·408	0·836	0·980
	r_M	**0·049**	0·039	0·034	0·017	0·009
2	r_P	0·148	0·186	0·270	0·265	0·861
	r_M	0·054	**0·049**	**0·048**	0·033	0·032
3	r_P	0·123	0·159	0·206	0·494	0·745
	r_M	0·054	0·053	0·048	**0·046**	**0·044**
5	r_P	0·102	0·123	0·164	0·349	0·551
	r_M	0·054	0·053	0·053	0·041	0·046
200	r_P	**0·055**	**0·058**	0·066	0·080	0·095
	r_M	0·051	0·053	0·055	0·052	0·057
500	r_P	0·053	0·056	**0·056**	**0·060**	0·074
	r_M	0·050	0·050	0·050	0·050	0·052
2000	r_P	0·051	0·049	0·053	0·047	**0·060**
	r_M	0·049	0·047	0·050	0·050	0·052

Table 3 reports coverage probabilities of the 0·05 quantile of the standard normal distribution for r_P and r_M. Similar results were found in the other tail. The parameter of interest is $\psi = 1$. Note that the accuracy of $r_M(\psi)$ is affected little by large numbers of nuisance parameters, whereas the accuracy of $r_P(\psi)$ tends to degenerate for moderately large values of q.

Example 5: *Loblolly data.* We consider a dataset concerning the growth of Loblolly pine trees; see Appendix A.13 of Pinheiro & Bates (2000). The data consist of $q = 14$ trees with different seed sources and for each of them we have $m = 6$ observations for the height (in feet), y, with respect to 6 different ages (in years) of the tree, namely $x \in \{3, 5, 10, 15, 20, 25\}$. We assume that $Y_{ij} = \mu(\beta_i;\, x_j) + \sigma\varepsilon_{ij}$, where the ε_{ij} are independent standard normal random variables and

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

Here we consider $\psi = \sigma^2$ and $\lambda_i = \beta_i$. The maximum likelihood estimates $\hat{\beta}_i$ of the β_i's are obtained using nonlinear least squares. The maximum likelihood estimator of σ^2 is

$$\hat{\sigma}^2 = (mq)^{-1} \sum_{i=1}^{q} \sum_{j=1}^{m} \{y_{ij} - \mu(\hat{\beta}_i;\, x_j)\}^2$$

and the profile loglikelihood is $l_P(\sigma^2) = -\frac{1}{2}mq(\log \sigma^2 + \hat{\sigma}^2/\sigma^2)$. Since $\hat{\beta}_i = \hat{\beta}_{i\sigma^2}$, the modified profile loglikelihood coincides with the Cox & Reid (1987) adjusted profile loglikelihood and is $l_M(\sigma^2) = l_P(\sigma^2) + \frac{3}{2}q \log \sigma^2$. This gives the estimator $\hat{\sigma}^2_M = m\hat{\sigma}^2/(m-3) = 2\hat{\sigma}^2$. In this

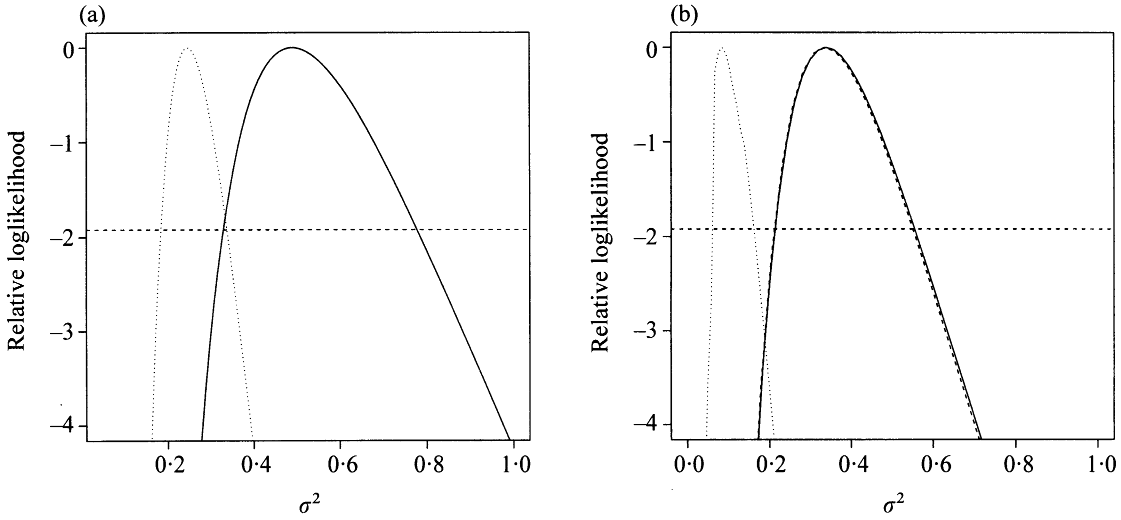

Fig. 2: Example 5. Loblolly data. Relative loglikelihoods for σ^2 when the distribution is (a) normal and (b) t_3: profile (dotted), modified profile (solid), Cox & Reid (1987) (dashed). The horizontal dashed line gives the 0·95 confidence interval.

case, we have $\hat\sigma^2 = 0{\cdot}2453$ and $\hat\sigma^2_{\mathrm{M}} = 0{\cdot}4906$. Figure 2(a) shows the relative loglikelihoods for $\hat\sigma^2$ together with the 0·95 confidence intervals based on the asymptotic distribution of the likelihood ratio statistics.

A simulation with 10 000 replications, $\beta_i = \hat\beta_i$ and $\sigma^2 = 0{\cdot}5$ gave coverage probabilities for the nominal 0·95 confidence interval equal to 0·025 for the profile likelihood and 0·951 for the modified profile likelihood.

Suppose now that ε_{ij} are independent random variables with the Student distribution t_ν, with $\nu = 3$ degrees of freedom. In this case, $\hat\beta_{i\sigma^2}$ is no longer equal to $\hat\beta_i$, even though β_i is still orthogonal to σ^2. The modified profile likelihood may be computed using the approximation of sample space derivatives given in § 9.5.3 of Severini (2000), due to Fraser & Reid (1995). In particular, using $\{y_{ij} - \mu(\beta_i;\, x_j)\}/\sigma$ as pivotal quantity, we have that the approximation $\tilde l_{\beta_i;\hat\beta_i}(\sigma^2, \beta_i)$, which is a 3×3 matrix, is proportional to

$$\sum_{j=1}^{m} \frac{\nu\sigma^2 - \{y_{ij} - \mu(\beta_i;\, x_j)\}^2}{[\nu\sigma^2 + \{y_{ij} - \mu(\beta_i;\, x_j)\}^2]^2}\, \mu'(\beta_i;\, x_j)\mu'(\hat\beta_i;\, x_j)^{\mathrm{T}},$$

where $\mu'(\beta_i;\, x_j) = \partial\mu(\beta_i;\, x_j)/\partial\beta_i$ is a 3×1 vector. In this case, the estimates are $\hat\sigma^2 = 0{\cdot}0833$ and $\hat\sigma^2_{\mathrm{M}} = 0{\cdot}3396$. Figure 2(b) shows the relative loglikelihoods.

6. Final remarks

Remark 1. Sometimes modified profile likelihoods perform better than the profile likelihood for reasons other than the score bias, which usually leads to a location adjustment. Even though this paper does not deal with that issue, modified profile likelihoods may also give a sensible spread adjustment; see DiCiccio et al. (1996). For example, suppose that Y_{ij} are independent normal random variables with mean μ and variances σ_i^2, as in Example 9.18 in Severini (2000). When $m_i = m$ we have $\hat\mu = \hat\mu_{\mathrm{M}}$, but the confidence intervals based on W_{P} tend to be too narrow. For instance, in a simulation with 10 000 replications, $m = 5$ and $q = 25$, we obtained actual coverages of 0·866 for W_{P} and 0·947 for W_{M}, when the nominal level was 0·95.

Remark 2. When ψ is a scalar, there is a modified directed likelihood (Barndorff-Nielsen 1986; Barndorff-Nielsen & Cox, § 6.6; Fraser et al., 1999), involving more than the modified profile likelihood, that is useful for computing p-values and confidence limits. This provides very accurate results even in rather extreme situations. Pierce & Peters (1992) show that the modified directed likelihood can be written as $r_{\mathrm{P}} + \mathrm{NP} + \mathrm{INF}$, where NP is a nuisance parameter adjustment and INF is an information adjustment. Both these adjustments are of order $O_p(n^{-1/2})$ and the modified profile likelihood essentially includes the NP adjustment. In fact, $r_{\mathrm{M}} = r_{\mathrm{P}} + \mathrm{NP} + o_p(1)$ (Sartori et al., 1999). In $(m \times q)$-asymptotics, from the results in this paper together with some other simple expansions, it is possible to see that $\mathrm{NP} = O_p(qn^{-1/2})$, while INF has the usual order $O_p(n^{-1/2})$. This is true regardless of the nature of the sequence $\{q, m\}$. Moreover, it can be shown that the relationship $r_{\mathrm{M}} = r_{\mathrm{P}} + \mathrm{NP} + o_p(1)$ is still true, provided that $1/m = o(q^{-1/3})$. This means that, under this condition, r_{M} tends to be very close to the modified directed likelihood; see also Sartori et al. (1999) for some simulation results.

Remark 3. Severini (2002) has investigated modified profile estimating functions for partially specified models where a likelihood function is not available. Under his conditions, the results of § 4 about $\hat{\psi}$ and $\hat{\psi}_{\mathrm{M}}$ extend to profile and modified profile estimating functions.

Appendix

Some technical details

Consider the central limit theorem for independent variables, where X_i are independent random variables with means μ_i and variances σ_i^2 $(i = 1, \ldots, q)$. Then, under certain conditions, such as those of Lyapounov's theorem, we have that

$$T_q = \frac{\sum_{i=1}^{q}(X_i - \mu_i)}{C_q} \rightarrow N(0, 1),$$

in distribution, where $C_q = (\sum_{i=1}^{q} \sigma_i^2)^{1/2}$. This implies that $T_q = O_p(1)$ and

$$\sum_{i=1}^{q} X_i = O\left(\sum_{i=1}^{q} \mu_i\right) + O_p(C_q). \tag{A1}$$

We use this result to evaluate the order of terms in (6), assuming uniformity over strata of the error bounds of the single terms in the sums. The term B_{P} is of order $O_p(q)$ since $E_\theta(B_{\mathrm{P}}^i) = O(1)$ and $V_\theta(B_{\mathrm{P}}^i) = O(1)$. This implies that in (A1) we have $\sum_{i=1}^{q} \mu_i = O(q)$ and $C_q = O(q^{1/2})$, giving $B_{\mathrm{P}} = O(q) + O_p(q^{1/2}) = O_p(q)$. The term R_{P} is such that $E_\theta(R_{\mathrm{P}}^i) = O(m^{-1})$ and $V_\theta(R_{\mathrm{P}}^i) = O(m^{-1})$. This implies that $R_{\mathrm{P}} = O(q/m) + O_p\{(q/m)^{1/2}\} = O_p[\max\{q/m, (q/m)^{1/2}\}]$.

The modified profile score function can be written in the form

$$U_{\mathrm{M}} = U_{\psi|\lambda} + B_{\mathrm{M}} + R_{\mathrm{P}},$$

where $B_{\mathrm{M}} = B_{\mathrm{P}} + \partial M/\partial\psi = \sum_{i=1}^{q}(B_{\mathrm{P}}^i + \partial M^i/\partial\psi) = \sum_{i=1}^{q} B_{\mathrm{M}}^i$. This follows from (6) and (4). From (5), it is straightforward to see that $E_\theta(B_{\mathrm{M}}^i) = O(m^{-1})$ and, if we use the delta method, that $V_\theta(B_{\mathrm{M}}^i) = O(m^{-1})$. This means that B_{M} and R_{P} are quantities of the same order. Hence, we can write $U_{\mathrm{M}} = U_{\psi|\lambda} + R_{\mathrm{M}}$, where

$$R_{\mathrm{M}} = B_{\mathrm{M}} + R_{\mathrm{P}} = O_p[\max\{q/m, (q/m)^{1/2}\}].$$

Finally, we consider the relationship between the partial information and the observed profile information $j_{\mathrm{P}}(\psi) = \sum_{i=1}^{q} j_{\mathrm{P}}^i(\psi)$,where

$$j_{\mathrm{P}}^i(\psi) = j_{\psi\psi}^i(\psi, \hat{\lambda}_{i\psi}) - j_{\lambda_i\lambda_i}^{-1}(\psi, \hat{\lambda}_{i\psi}) j_{\psi\lambda_i}^2(\psi, \hat{\lambda}_{i\psi})$$

is the observed profile information in the ith stratum. Expanding each single term around (ψ, λ), using an expansion for $(\hat{\lambda}_{i\psi} - \lambda_i)$ as in formula (9.87) in Pace & Salvan (1997), we have that

$$j_{\mathrm{P}}(\psi) = i_{\psi\psi|\lambda} + O_p(q) + O_p(n^{1/2}). \tag{A2}$$

Moreover, since $i_{\psi\psi|\lambda} = n\bar{\imath}_{\psi\psi|\lambda}$, we can write

$$j_{\mathrm{P}}(\psi) = i_{\psi\psi|\lambda}\{1 + O_p(m^{-1}) + O_p(n^{-1/2})\} = i_{\psi\psi|\lambda}\{1 + o_p(1)\}. \tag{A3}$$

The relationship between the observed information and the partial information is the same even for a modified profile likelihood. In fact, since derivatives of $M(\psi)$ are of order $O_p(q)$, we have

$$j_{\mathrm{M}}(\psi) = j_{\mathrm{P}}(\psi) - \partial^2 M(\psi)/\partial\psi^2 = j_{\mathrm{P}}(\psi) + O_p(q) = i_{\psi\psi|\lambda} + O_p(q) + O_p(n^{1/2}),$$

where the last step follows from (A2).

Acknowledgement

A substantial part of this work has been done while I was visiting the Departments of Statistics of the University of Toronto and of the Northwestern University. I would like to thank both departments for their kind hospitality and, in particular, Don Fraser, Nancy Reid and Tom Severini for interesting discussions. I am most grateful to Alessandra Salvan, Luigi Pace and Ruggero Bellio for their help and encouragement. Finally, I would like to thank an associate editor for insightful comments, which significantly improved the paper. This research has been partially supported by a grant from Ministero dell'Istruzione, dell'Università e della Ricerca, Italy.

References

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *J. R. Statist. Soc.* B **32**, 283–301.

Andersen, E. B. (1971). Asymptotic properties of conditional likelihood ratio tests. *J. Am. Statist. Assoc.* **66**, 630–3.

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70**, 343–65.

Barndorff-Nielsen, O. E. (1986). Inference on full and partial parameters based on standardized signed log likelihood ratio. *Biometrika* **73**, 307–22.

Barndorff-Nielsen, O. E. (1994). Adjusted versions of profile likelihood and directed likelihood and extended likelihood. *J. R. Statist. Soc.* B **56**, 125–40.

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

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

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

Cox, D. R. & Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with Discussion). *J. R. Statist. Soc.* B **49**, 1–39.

Cox, D. R. & Reid, N. (1992). A note on the difference between profile and modified profile likelihood. *Biometrika* **79**, 408–11.

Davison, A. C. (1988). Approximate conditional inference in generalized linear models. *J. R. Statist. Soc.* B **50**, 445–61.

Davison, A. C. (1992). Discussion of the paper by D. A. Pierce & D. Peters (1992). *J. R. Statist. Soc.* B **54**, 726–7.

DiCiccio, T. J., Martin, M. A., Stern, S. E. & Young, G. A. (1996). Information bias and adjusted profile likelihoods. *J. R. Statist. Soc.* B **58**, 189–203.

Fraser, D. A. S. & Reid, N. (1995). Ancillaries and third-order significance. *Utilitas Math.* **47**, 33–53.

Fraser, D. A. S., Reid, N. & Wu, J. (1999). A simple and general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86**, 249–64.

McCullagh, P. & Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *J. R. Statist. Soc.* B **52**, 325–44.

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

Pace, L. & Salvan, A. (1997). *Principles of Statistical Inference from a Neo-Fisherian Perspective.* Singapore: World Scientific Publishing Co.

Pierce, D. A. & Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families (with Discussion). *J. R. Statist. Soc.* B **54**, 701–37.

Pinheiro, J. C. & Bates, D. M. (2000). *Mixed-effects Models in S and S-PLUS.* New York: Springer-Verlag.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16**, 356–66.

Sartori, N., Bellio, R., Salvan, A. & Pace, L. (1999). The directed modified profile likelihood in models with many nuisance parameters. *Biometrika* **86**, 735–42.

Severini, T. A. (1998). An approximation to the modified profile likelihood function. *Biometrika* **85**, 403–11.

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

Severini, T. A. (2002). Modified estimating functions. *Biometrika* **89**, 333–43.

[*Received April* 2001. *Revised October* 2002]