A Simple Method for the Adjustment of Profile Likelihoods

Peter McCullagh; Robert Tibshirani

Journal of the Royal Statistical Society. Series B (Methodological), Volume 52, Issue 2 (1990), 325-344.

Stable URL:
http://links.jstor.org/sici?sici=0035-9246%281990%2952%3A2%3C325%3AASMFTA%3E2.0.CO%3B2-7

J. R. Statist. Soc. B (1990)
52, *No.* 2, *pp.* 325–344

A Simple Method for the Adjustment of Profile Likelihoods

By PETER McCULLAGH and ROBERT TIBSHIRANI†

University of Chicago, USA

University of Toronto, Canada

[Received July 1988. Final revision May 1989]

SUMMARY

We propose a simple adjustment for profile likelihoods. The aim of the adjustment is to alleviate some of the problems inherent in the use of profile likelihoods, such as bias, inconsistency and overoptimistic variance estimates. The adjustment is applied to the profile log-likelihood score function at each parameter value so that its mean is zero and its variance is the negative expected derivative matrix of the adjusted score function. For cases in which explicit calculation of the adjustments is difficult, we give two methods to simplify their computation: an 'automatic' simulation method that requires as input only the profile log-likelihood and its first few derivatives; first-order asymptotic expressions. Some examples are provided and a comparison is made with the conditional profile log-likelihood of Cox and Reid.

Keywords: ASYMPTOTIC THEORY; BOOTSTRAP; NUISANCE PARAMETERS; PROFILE LIKELIHOOD; SCORE FUNCTION

1. INTRODUCTION

Inference in the presence of nuisance parameters is a widely encountered and difficult problem, particularly for a frequency-based theory of inference. Probably the simplest approach is to maximize out the nuisance parameters for fixed values of the parameters of interest and to construct the so-called profile likelihood. The profile likelihood is then treated as an ordinary likelihood function for estimation and inference about the parameters of interest. Unfortunately, with large numbers of nuisance parameters, this procedure can produce inefficient or even inconsistent estimates. A simple example in which these phenomena occur is the 'many normal means' problem where we observe Y_{i1}, Y_{i2}, each independent and normally distributed with means μ_i and variance σ^2, for $i = 1, 2, \ldots, n$. The variance σ^2 is taken as the parameter of interest. In this case the maximizer of the profile likelihood (also the global maximum likelihood estimate) $\hat{\sigma}^2$ has expectation $\sigma^2/2$ and is inconsistent. The reason is that the bias of the usual maximum likelihood estimate of variance accumulates across the data pairs.

An integrated likelihood (Kalbfleisch and Sprott, 1970) can be obtained if one is willing to specify a joint prior distribution for the parameters, or at least a conditional prior distribution for the nuisance parameters given the parameters of interest. In the absence of prior information, however, this is difficult to do in an 'objective' way.

In some special problems, marginal or conditional likelihoods can be constructed from which valid inferences can be made (see, for example, Cox and Hinkley (1974)). For example, in the many normal means problem, a marginal likelihood based on the

†*Address for correspondence*: Department of Preventive Medicine and Biostatistics, University of Toronto, Toronto, Ontario, M5S 1A8, Canada.

0035–9246/90/52325 $2.00

sample variances can be derived, and this likelihood has its maximum at $2\hat{\sigma}^2$, the unbiased estimate of σ^2. Such problems are relatively rare, however, and hence there is a need for a more general approach. Two recent advances in this area are the modified profile likelihood (Barndorff-Nielsen, 1986) and the closely related conditional profile likelihood (Cox and Reid, 1987). Both of these modifications attempt to adjust the profile likelihood for nuisance parameters. They correct the inconsistency of the profile likelihood in some problems (like the problem above) and automatically make 'degrees of freedom' adjustments in normal theory cases where accepted answers are available for comparison. As a result, the likelihood ratio statistic derived from the modified or conditional profile likelihood seems to be more nearly approximated by a χ_1^2 distribution (when there is a single parameter of interest) than is that derived from the profile likelihood. The construction of these likelihoods beyond simple cases is an open question, however.

In this paper we propose an alternative simpler approach to this problem. Our goal is to adjust the profile log-likelihood so that the mean of the score function is zero and the variance of the score function equals its negative expected derivative matrix. In the terminology of Godambe (1960) and Lindsay (1982), our goal is to adjust the profile log-likelihood score function so that it is unbiased and information unbiased. The hope is that, by making these adjustments, the asymptotic behaviour of the quantities derived from the likelihood (e.g. its maximizer, information matrix and confidence sets) will be improved. However, as we state in Section 6, we have no strong argument to support this claim. In addition, like the profile likelihood, our adjusted profile likelihood is not a likelihood in the usual sense and hence does not correspond to densities of observable events.

We discuss both exact and approximate methods for the calculation of the adjustments. The exact calculation is achieved, in principle at least, through a simulation or 'parametric bootstrap' process in which the moments of the profile log-likelihood score function are estimated by parametric bootstrap sampling. The moments are computed at each value of the parameter of interest, with the restricted maximum likelihood estimate used for the nuisance parameters. The estimated moments are then used to centre and rescale the profile log-likelihood score function. The approximate adjustment uses first-order asymptotic expressions for the cumulants of the derivatives of the profile log-likelihood score function.

The adjusted log-likelihood is given by the integral of the adjusted score function. The result, which we call the 'adjusted profile log-likelihood' ($l_{\mathrm{ap}}(\psi)$), is parameterization invariant and, in the examples that we discuss, seems to correct the profile log-likelihood in a similar manner to the modified and conditional profile log-likelihoods. All that is required for the exact computation of l_{ap} is a routine for calculating the profile log-likelihood and its first two derivatives, and a routine for sampling from the estimated probability model. Thus the method can be applied to complex models in an 'automatic' way. However, of the order of 1000 profile log-likelihoods need to be computed and hence the workload could be prohibitive in some problems. In addition, if no explicit form exists for the profile log-likelihood (and this is not uncommon) the computational challenge might be substantial. We have not yet tackled such a case.

In Section 2 we introduce notation and define the conditional and modified profile log-likelihoods. Section 3 defines the adjusted profile log-likelihood and gives the bootstrap algorithm for its estimation. Section 4 discusses the first-order approximations and Section 5 contains some examples. In Section 6 we discuss some heuristic justifications for the procedure.

2. PROFILE LIKELIHOOD AND SOME MODIFICATIONS

In this section we establish some notation and review the modified and conditional profile likelihoods. We begin with a random n sample of d vectors $\mathcal{Y}=(Y_1, \ldots, Y_n)$, each Y_i independently and identically distributed with density $f_Y(y, \theta)$. The parameter θ can be partitioned as $\theta=(\psi, \lambda)$ where $\psi=(\psi_1, \ldots, \psi_r)$ is the parameter of interest and $\lambda=(\lambda_1, \ldots, \lambda_p)$ is the nuisance parameter. The log-likelihood is denoted by $l(\theta)$.

Denote by $\hat{\theta}=(\hat{\psi}, \hat{\lambda})$ the overall maximum likelihood estimate. Let $\hat{\psi}_\lambda$ denote the maximum likelihood estimate of ψ for fixed λ and similarly $\hat{\lambda}_\psi$. The profile log-likelihood for ψ is defined by

$$l_p(\psi)=l(\psi, \hat{\lambda}_\psi).$$

Log-likelihoods are defined only up to additive functions of the data, but since we shall be dealing solely with its derivatives there is no need to make this definition more precise.

For the single parameter of interest case ($r=1$) the conditional profile log-likelihood (Cox and Reid, 1987) is defined by

$$l_{cp}(\psi)=l(\psi, \hat{\lambda}_\psi)-\tfrac{1}{2}\log\{\det nj_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\}$$

where $j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)$ is the observed information per observation for the λ components. This definition requires that ψ and λ be orthogonal in the sense defined by Cox and Reid, i.e. $E(-\partial^2 l/\partial\psi\, \partial\lambda_j)=0$ for $j=1, 2, \ldots, p$. This is one of several similar definitions given by Cox and Reid; it corresponds to their expression (10) and is probably the most easily computable definition of all. The interpretation of the correction term in $l_{cp}(\psi)$ is that it penalizes values of ψ for which the information about λ is relatively large. The derivation of $l_{cp}(\psi)$ uses a double-conditioning argument as well as several approximations with error $O_p(n^{-1})$, including Barndorff-Nielsen's (1983) formula for the distribution of the maximum likelihood estimator. A simple alternative derivation can be derived via Bayesian arguments (Sweeting, 1987; R. Kass, personal communication). Given orthogonal parameters ψ and λ, if independent priors are assumed, then the ratio of the marginal posterior for ψ to the prior for ψ equals $\exp l_{cp}(\psi)$ to order n^{-1}. This result holds regardless of the prior assumed for λ.

The modified profile log-likelihood (Barndorff-Nielsen, 1986) is defined by

$$l_{mp}(\psi)=l(\psi, \hat{\lambda}_\psi)-\tfrac{1}{2}\log\{\det nj_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\}+\log\{\det(\mathrm{d}\hat{\lambda}_\psi/\mathrm{d}\hat{\lambda})\}$$

This definition does not require the orthogonality of ψ and λ; the last term on the right-hand side can be thought of as a correction for non-orthogonality.

3. SIMPLE ADJUSTMENT

Denote the score function derived from the profile log-likelihood by

$$U(\psi)=\frac{\partial}{\partial\psi}l_p(\psi).$$

A basic property of regular maximum likelihood score functions is that their mean is zero, and their variance is minus their expected derivative matrix, expectations being computed at the true parameter value. Since we are interested in the profile log-likelihood, we seek to adjust $U(\psi)$ so that these properties hold when expectations and derivatives are computed at $(\psi, \hat{\lambda}_\psi)$ rather than at the true parameter point. For simplicity we consider the single parameter of interest ($r = 1$) case here; we give details of the general case later in this section.

Consider functions $m(\psi)$ and $w(\psi)$ and let

$$\tilde{U}(\psi) = \{U(\psi) - m(\psi)\}\, w(\psi).$$

Then we require

$$E_{\psi, \hat{\lambda}_\psi}\, \tilde{U}(\psi) = 0 \tag{1}$$

$$\operatorname{var}_{\psi, \hat{\lambda}_\psi}(\tilde{U}(\psi)) = -E_{\psi, \hat{\lambda}_\psi} \frac{\partial}{\partial\psi}\, \tilde{U}(\psi) \tag{2}$$

for all ψ, the subscripts indicating that expectations are computed under $(\psi, \hat{\lambda}_\psi)$. Solving for $m(\psi)$ and $w(\psi)$, we find

$$m(\psi) = E_{\psi, \hat{\lambda}_\psi}\, U(\psi)$$

$$w(\psi) = \left\{-E_{\psi, \hat{\lambda}_\psi} \frac{\partial^2}{\partial\psi^2}\, l_\mathrm{p}(\psi) + \frac{\partial}{\partial\psi}\, m(\psi)\right\} \Big/ \operatorname{var}_{\psi, \hat{\lambda}_\psi}(U(\psi)).$$

Finally, let

$$l_\mathrm{ap}(\psi) = \int^{\psi} \tilde{U}(t)\, \mathrm{d}t, \tag{3}$$

the 'adjusted profile log-likelihood' for ψ. The exponential of this function will be called the 'adjusted profile likelihood'.

The required ingredients for the computation of $l_\mathrm{ap}(\psi)$ are

$$m(\psi) = E_{\psi, \hat{\lambda}_\psi}\, U(\psi), \qquad \operatorname{var}_{\psi, \hat{\lambda}_\psi}(U(\psi)), \qquad E_{\psi, \hat{\lambda}_\psi} \frac{\partial^2}{\partial\psi^2}\, l_\mathrm{p}(\psi) \quad \text{and} \quad \frac{\partial}{\partial\psi}\, m(\psi).$$

All these involve expectations with respect to $f_Y(y, (\psi, \hat{\lambda}_\psi))$. Sometimes these expectations can be computed analytically, as in several of the simple examples given later, but in general we must resort to Monte Carlo simulation. The algorithm for computing $l_\mathrm{ap}(\psi)$ is as follows.

(a) Compute $l_\mathrm{p}(\psi)$ and $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$.

(b) For each value ψ over a grid

  (i) sample B times from $f_{Y^*}(y, \hat{\theta}_\psi)$; let $l_j^*(\psi)$ and $u_j^*(\psi)$ be the profile log-likelihood and score function from the jth bootstrap sample,

  (ii) let

$$m(\psi) = \sum_1^B u_j^*(\psi)/B, \qquad \widehat{\frac{\partial^2}{\partial\psi^2}\, l_\mathrm{p}(\psi)} = \sum_1^B \frac{\partial^2}{\partial\psi^2}\, l_j^*(\psi)/B,$$

$$w(\psi) = \left\{-\widehat{\frac{\partial^2}{\partial\psi^2}\, l_\mathrm{p}(\psi)} + \frac{\partial}{\partial\psi}\, m(\psi)\right\} \Big/ \sum_1^B \frac{\{u_j^*(\psi) - m(\psi)\}^2}{B}.$$

(c) Set

$$l_{\mathrm{ap}}(\psi) = \int^{\psi} \{U(t) - m(t)\} w(t)\, \mathrm{d}t.$$

In the next section we derive a first-order approximation to $m(\psi)$. This approximation is reasonably simple and may be satisfactorily accurate in many problems.

The problem of setting confidence limits for ψ, treating λ as a nuisance parameter, is invariant under non-singular transformations that preserve ψ, namely $\theta \rightarrow \theta' = (\psi', \lambda')$, where

$$\psi' = \psi, \qquad \lambda' = g(\lambda, \psi), \tag{4}$$

where $g(\cdot, \psi)$ is invertible for each fixed ψ. The profile log-likelihood $l_{\mathrm{p}}(\psi)$ is also invariant under this group of transformations. Hence the adjusted profile log-likelihood is invariant. The main implication of this result is that confidence regions based on $l_{\mathrm{p}}(\psi)$ are unaffected by the parameterization of λ, in contrast with equation (8) of Cox and Reid (1987) which requires orthogonality. However, orthogonality is often convenient for the computations, as will be seen in some of the examples.

If $m(\psi) = 0$, the maximizer $\tilde{\psi}$ of $l_{\mathrm{ap}}(\psi)$ equals the maximum likelihood estimate $\hat{\psi}$, i.e. the procedure may reshape but does not shift $l_{\mathrm{p}}(\psi)$ when $E_{\psi, \lambda_\psi} U(\psi) = 0$ for all ψ. In contrast, if $m(\psi) \neq 0$ then $l_{\mathrm{ap}}(\psi)$ may differ in location from $l_{\mathrm{p}}(\psi)$ even if $w(\psi) = 1$.

If ψ has $r > 1$ components, then the adjusted profile log-likelihood can be generalized in a straightforward manner although its computation will be more costly. We take $\tilde{U}(\psi) = w(\psi)\{U(\psi) - m(\psi)\}$ where $\tilde{U}(\psi)$, $U(\psi)$ and $m(\psi)$ are column vectors $((U(\psi))_j = (\partial/\partial\psi_j) l_{\mathrm{p}}(\psi))$, and $w(\psi)$ is an $r \times r$ matrix. The conditions determining $m(\psi)$ and $w(\psi)$ are the same as equations (1) and (2), except that they now involve $r \times 1$ vectors and $r \times r$ matrices respectively. The solutions are

$$m(\psi) = E_{\psi, \lambda_\psi} U(\psi)$$

$$w(\psi) = \{\mathrm{var}_{\psi, \lambda_\psi}(U(\psi))\}^{-1} \left\{ -E_{\psi, \lambda_\psi} \frac{\partial^2}{\partial\psi^2} l_{\mathrm{p}}(\psi) + \frac{\partial}{\partial\psi} m(\psi) \right\}^{\mathrm{T}}.$$

To compute these in general we would need to simulate over a grid in r-dimensional space.

One additional difficulty that occurs only in the multiparameter case is that in general $\partial\tilde{U}_r/\partial\psi^s \neq \partial\tilde{U}_s/\partial\psi^r$. Consequently, in general there is no function $l_{\mathrm{ap}}(\psi)$ having gradient vector $\tilde{U}_r$.

Remark 1. Our present implementation of the algorithm treats only the $r = 1$ case. We use a grid of 40 equally spaced ψ values and $B = 25$ bootstrap samples at each of these values. First differences are used to estimate the derivative of $m(\psi)$, and a trapezoid rule is used to estimate the final integral. Programming was done in the new S language (Becker *et al.*, 1988). A typical computation of $l_{\mathrm{ap}}(\psi)$ took about 20 s on a SUN 3/160 computer. A Fortran implementation would run considerably faster.

Remark 2. It is interesting to examine why the expectation of the profile log-likelihood is not zero in general, and similarly why the variance of the score does not equal minus the expected derivative of the profile log-likelihood score. Consider a log-likelihood

$$l(\theta) = \sum_1^n \log f_Y(y_i, \theta).$$

Let

$$f_{\mathscr{Y}}(y, \theta) = \prod_1^n f_Y(y_i, \theta).$$

Then the usual proof of $E_\theta \partial l(\theta)/\partial\theta = 0$ goes as follows (see, for example, Silvey (1975)).

$$\begin{aligned} E_\theta \frac{\partial l(\theta)}{\partial\theta} &= \int \frac{\partial l(\theta)}{\partial\theta} f_{\mathscr{Y}}(y, \theta)\, \mathrm{d}y \\ &= \int \frac{(\partial/\partial\theta) f_{\mathscr{Y}}(y, \theta)}{f_{\mathscr{Y}}(y, \theta)} f_{\mathscr{Y}}(y, \theta)\, \mathrm{d}y \\ &= \int \frac{\partial}{\partial\theta} f_{\mathscr{Y}}(y, \theta)\, \mathrm{d}y \\ &= \frac{\partial}{\partial\theta} \int f_{\mathscr{Y}}(y, \theta)\, \mathrm{d}y \\ &= \frac{\partial}{\partial\theta} 1 \\ &= 0. \end{aligned}$$

Sufficient regularity is assumed to allow the interchange in the second to last step. Now if $l(\theta)$ is instead a profile log-likelihood, i.e. $l(\theta) = l(\psi, \hat{\lambda}_\psi)$, then the expression corresponding to the second line of the proof is

$$\int \frac{(\partial/\partial\psi) f_{\mathscr{Y}}(y^*, (\psi, \hat{\lambda}_\psi^*))}{f_{\mathscr{Y}}(y^*, (\psi, \hat{\lambda}_\psi^*))} f_{\mathscr{Y}}(y^*, (\psi, \hat{\lambda}_\psi))\, \mathrm{d}y^*$$

where we have introduced the asterisk notation to indicate quantities that involve the arguments of integration y_i^*. As can be seen, the necessary cancellation does not occur. The point is that $\hat{\lambda}_\psi^*$ is a function of the data and this brings in a new source of randomness into the expectation. A similar phenomenon occurs in the usual proof of $\mathrm{var}_\theta(\partial l(\theta)/\partial\theta) = -E_\theta(\partial^2 l(\theta)/\partial\theta^2)$.

3.1. *Example 1: Many Normal Means*

In this problem we observe n data pairs (Y_{i1}, Y_{i2}) with each Y_{ij} independently distributed $N(\mu_i, \sigma^2)$, σ^2 being the parameter of interest. The profile log-likelihood is

$$l_{\mathrm{p}}(\sigma^2) = -n \log \sigma^2 - s^2/2\sigma^2$$

where $s^2 = \Sigma_1^n (y_{i1} - y_{i2})^2/2$. The maximum of $l_{\mathrm{p}}(\sigma^2)$ is $s^2/2n$, an inconsistent estimate with expectation $\sigma^2/2$. The modified profile log-likelihood is easily derived as

$$l_{\mathrm{mp}}(\sigma^2) = -(n/2) \log \sigma^2 - s^2/2\sigma^2$$

which is identical with the conditional profile log-likelihood. The marginal log-likelihood of s^2 also equals $l_{\mathrm{mp}}(\sigma^2)$. The maximum of $l_{\mathrm{mp}}(\sigma^2)$ is s^2/n which is unbiased and consistent.

Because of the simple nature of this problem, we can carry out the exact adjustments exactly. The adjustments turn out to be $m(\sigma^2) = -n/2\sigma^2$, $w(\sigma^2) = 1$ and thus $l_{\mathrm{ap}}(\sigma^2)$ agrees with the marginal log-likelihood as well.

4. FIRST-ORDER APPROXIMATION FOR ADJUSTMENTS

The derivatives of $l_{\mathrm{p}}(\psi)$ may be written in terms of the derivatives of $l(\psi, \lambda)$ at the true parameter point as follows.

$$\begin{aligned}\frac{\partial l_{\mathrm{p}}}{\partial \psi} &= \frac{\partial}{\partial \psi}\, l(\psi, \hat{\lambda}_\psi) \\ &= \frac{\partial l}{\partial \psi} + \frac{\partial^2 l}{\partial \psi\, \partial \lambda}\,(\hat{\lambda}_\psi - \lambda) + \frac{1}{2}\,\frac{\partial^3 l}{\partial \psi\, \partial \lambda^2}\,(\hat{\lambda}_\psi - \lambda)^2 + \ldots.\end{aligned} \tag{5}$$

Under the usual regularity conditions in which the joint maximum likelihood estimate is consistent, the first three terms in this expansion are $O_{\mathrm{p}}(n^{1/2})$, $O_{\mathrm{p}}(n^{1/2})$ and $O_{\mathrm{p}}(1)$. The remainder terms are $O_{\mathrm{p}}(n^{-1/2})$. The first term has mean zero but the remaining two terms have expectation $O(1)$ under the same regularity conditions.

In calculating the bias correction it is most convenient to use index notation explicitly. The components of ψ are denoted by ψ^r and the derivatives of l with respect to ψ are denoted by

$$U_r = \frac{\partial l}{\partial \psi^r}, \qquad U_{rs} = \frac{\partial^2 l}{\partial \psi^r\, \partial \psi^s}.$$

The components of λ are denoted by $\lambda^i, \lambda^j, \ldots$, and the derivatives with respect to λ are

$$U_i = \frac{\partial l}{\partial \lambda^i}, \qquad U_{ij} = \frac{\partial^2 l}{\partial \lambda^i\, \partial \lambda^j}, \qquad U_{ri} = \frac{\partial^2 l}{\partial \psi^r\, \partial \lambda^i},$$

etc. For the cumulants of these derivatives we use the notation of McCullagh (1987), chapter 7. Thus

$$\begin{gathered}\kappa_r = E(U_r) = 0, \qquad \kappa_i = E(U_i) = 0, \\ \kappa_{rs} = E(U_{rs}) = -\kappa_{r,s}, \qquad \kappa_{ri} = E(U_{ri}) = -\operatorname{cov}(U_r, U_i), \\ \kappa_{r,ij} = \operatorname{cov}(U_r, U_{ij})\end{gathered}$$

etc. In addition $\kappa^{i,j}$ denotes the matrix inverse of $\kappa_{i,j}$.

From equation (7.10) of McCullagh (1987) we have

$$\hat{\lambda}^i_\psi - \lambda^i = \kappa^{i,j} U_j + \kappa^{i,j}\kappa^{k,l}(U_{jk} - \kappa_{jk}) U_l + \tfrac{1}{2}\kappa^{ijk} U_j U_k + O_{\mathrm{p}}(n^{-1}). \tag{6}$$

The sample size does not appear explicitly here but is incorporated into the random variables and the cumulants. Thus $\kappa_{i,j} = O(n)$, $\kappa^{i,j} = O(n^{-1})$ and

$$\kappa^{ijk} = \kappa^{i,i'}\kappa^{j,j'}\kappa^{k,k'}\kappa_{i'j'k'} = O(n^{-2}).$$

On substituting equation (6) into equation (5) and taking expectations, we find

$$E\left(\frac{\partial l_{\mathrm{p}}}{\partial \psi^r}\right) = -\tfrac{1}{2}(\kappa_{r,ij} - \kappa_{r,k}\kappa^{k,l}\kappa_{l,ij})\kappa^{i,j} - \tfrac{1}{2}(\kappa_{r,i,j} - \kappa_{r,k}\kappa^{k,l}\kappa_{l,i,j})\kappa^{i,j} + O(n^{-1}). \tag{7}$$

The leading terms are both $O(1)$. Details of the derivation are given in Appendix A.

The first-order bias correction (7) is remarkably simple. First it involves only the first derivative with respect to ψ and the first two derivatives with respect to λ. Intermediate calculations leading to equation (7) did involve third derivatives as in equation (6), but these were subsequently eliminated using Bartlett's identities. There is a formal similarity between equation (7) and the expression for the bias of the maximum likelihood estimate given in McCullagh (1987), p. 209, although the latter has no explicit recognition of nuisance parameters.

It is easy to give a simple description of the components of equation (7) using the following random variables:

$$\left.\begin{aligned} V_r &= U_r - \beta_r^i U_i = U_r - \kappa_{r,i}\kappa^{i,j} U_j, \\ Q_{11} &= U_i U_j \kappa^{i,j}, \\ Q_2 &= (U_{ij} - \kappa_{ij})\kappa^{i,j} = U_{ij}\kappa^{i,j} + \operatorname{rank}(\kappa^{i,j}). \end{aligned}\right\} \tag{8}$$

Thus V_r is the residual ψ derivative after linear regression on the λ derivatives; Q_{11} is the quadratic score statistic for λ and Q_2 is a linear function of the second derivatives with respect to λ. In terms of these we have

$$\begin{aligned} E\left(\frac{\partial l_\mathrm{p}}{\partial \psi^r}\right) &= -\tfrac{1}{2}\operatorname{cov}(V_r, Q_2) - \tfrac{1}{2}\operatorname{cov}(V_r, Q_{11}) + O(n^{-1}) \\ &= -\tfrac{1}{2}\operatorname{cov}(V_r, Q_{11} + Q_2) + O(n^{-1}). \end{aligned} \tag{9}$$

The expectations are computed at the true parameter point (ψ, λ) and hence the bias is a function of both parameters, not just ψ alone. In practice these quantities are computed at $(\psi, \hat{\lambda}_\psi)$.

Of the three statistics listed in equation (8) only V_r and Q_{11} are invariant with respect to mappings of the form (4): Q_2 is not invariant. However, $\operatorname{cov}(V_r, Q_2)$ is invariant, so that equation (9) is an invariant expression for the first-order bias.

A further simplification can be made in the special case of exponential family models. For the remainder of this section, we suppose that, for each fixed ψ, the log-likelihood for λ is a full exponential family model in which the dimension of the sufficient statistic equals the dimension of λ. In that case it is possible to reparameterize λ so that U_{ij} is a constant. The required parameterization is called the *canonical* parameterization. It follows then that $Q_2 = 0$. Hence the required first-order bias correction reduces to

$$E\left(\frac{\partial l_\mathrm{p}}{\partial \psi^r}\right) = -\tfrac{1}{2}\operatorname{cov}(V_r, Q_{11}). \tag{10}$$

It is not necessary here that the joint likelihood should be of the exponential family type: see example 8 later.

The variance adjustment is more complicated: we only give the first-order expression for a full exponential family. The covariance of the profile log-likelihood score is

$$\operatorname{cov}\left(\frac{\partial l_\mathrm{p}}{\partial \psi^r}, \frac{\partial l_\mathrm{p}}{\partial \psi^s}\right) = \kappa_{r,s} - \tfrac{1}{2}\kappa_{r,i,j}\kappa_{s,k,l}\kappa^{i,k}\kappa^{j,l} + O(n^{-1}). \tag{11}$$

The first term is $O(n)$ and the second term is $O(1)$. This derivation is similar to that for the mean adjustment and is not given here.

Now $\kappa_{r,s} = \mathrm{cov}(V_r, V_s)$ which equals the residual covariance of $\partial l/\partial\psi^r$ after linear regression on $(\partial l/\partial\lambda^i, \partial l/\partial\lambda^j, \ldots)$. Interestingly, expression (11) is the residual covariance of $\partial l/\partial\psi^r$ after *quadratic* regression on $(\partial l/\partial\lambda^i, \partial l/\partial\lambda^j, \ldots)$. Hence to this order

$$\mathrm{cov}\left(\frac{\partial l_{\mathrm{p}}}{\partial\psi_r}, \frac{\partial l_{\mathrm{p}}}{\partial\psi_s}\right) \leqslant \kappa_{r,s},$$

i.e. the difference is non-positive definite.

For a single parameter of interest and a single nuisance parameter we have

$$\kappa_{r,i,j} = E\left(\frac{\partial l_{\mathrm{p}}}{\partial\psi}\,\frac{\partial l_{\mathrm{p}}^2}{\partial\lambda}\right) - \beta E\left(\frac{\partial l_{\mathrm{p}}}{\partial\lambda}\right)^3.$$

The numerator of the variance correction is given by

$$-E\left\{\frac{\partial}{\partial\psi}\left(\frac{\partial l_{\mathrm{p}}}{\partial\psi^r} - b_r\right)\right\} = \kappa_{r,s} - \tfrac{1}{2}\kappa_{r,s,i}\kappa_{j,k,l}\kappa^{i,j}\kappa^{k,l} - \tfrac{1}{2}\kappa_{r,i,j}\kappa_{s,k,l}\kappa^{i,k}\kappa^{j,l} + O(n^{-1}) \qquad (12)$$

where $b_r = -\tfrac{1}{2}\kappa_{r,i,j}\kappa^{i,j}$, the (simplified) first-order bias expression for the full exponential family case. The first term in equation (12) is $O(n)$ and the remaining terms are $O(1)$. For a single parameter of interest, we can put all this together easily and we find that the adjusted score function has the form

$$\begin{aligned}\tilde{U}_r(\psi) &= \left(1 - \frac{\Delta_1}{\kappa_{r,s} - \Delta_2}\right)\left(\frac{\partial l_{\mathrm{p}}}{\partial\psi^r} - b_r\right) + O_{\mathrm{p}}(n^{-1}) \\ &= \frac{\partial l_{\mathrm{p}}}{\partial\psi^r} - b_r + \frac{\partial l_{\mathrm{p}}}{\partial\psi^r}\left(\frac{\Delta_1}{\kappa_{r,s} - \Delta_2}\right) + O_{\mathrm{p}}(n^{-1})\end{aligned} \qquad (13)$$

where $\Delta_1 = \tfrac{1}{2}\kappa_{r,s,i}\kappa_{j,k,l}\kappa^{i,j}\kappa^{k,l}$ and $\Delta_2 = \tfrac{1}{2}\kappa_{r,i,j}\kappa_{s,k,l}\kappa^{i,k}\kappa^{j,l}$.

5. FURTHER EXAMPLES

5.1. *Example 2: Normal Variance, Mean Unknown*

The problem of normal variance and mean unknown is similar to the many normal means problem, except that the differences between the various likelihoods disappear asymptotically because the number of nuisance parameters remains fixed. Given a sample $y_1, y_2, \ldots y_n$ from $N(\mu, \sigma^2)$, the profile log-likelihood is

$$l_{\mathrm{p}}(\sigma^2) = -(n/2)\log\sigma^2 - s^2/2\sigma^2$$

where $s^2 = \Sigma_1^n(y_i - \bar{y})^2$. Its maximum occurs at $\hat{\sigma}^2 = s^2/n$. The modified and conditional profile log-likelihoods replace the n in the first term by $n - 1$ and thus correct the bias in $\hat{\sigma}^2$. They are also equal to the marginal likelihood for s^2 and to the conditional log-likelihood given $\bar{y}$. It is easily shown that $l_{\mathrm{ap}}(\psi)$ equals $l_{\mathrm{cp}}(\sigma^2)$ and $l_{\mathrm{mp}}(\sigma^2)$ as in example 1. As an example of the invariance of l_{ap}, suppose that we use $\lambda = \mu + \sigma^2$ in place of the complementary (orthogonal) parameter μ. Then the profile log-likelihood $l_{\mathrm{p}}(\sigma^2)$ is unchanged, and the mean and variance of the profile log-likelihood score statistic, a function only of $s^2 \sim \sigma^2\chi_n^2$, are also unchanged.

The first-order approximations are straightforward to work out. Since $\partial l/\partial\sigma^2$ and $\partial l/\partial\mu$ are uncorrelated, we have $V = \partial l/\partial\sigma^2$ using the notation in equations (8). Further, $\partial^2 l/\partial\mu^2$ = constant, so that equation (10) gives

$$-\tfrac{1}{2}\,\mathrm{cov}(V, Q_{11}) = -1/2\sigma^2.$$

Also, the components of the variance correction in equation (13) are $\Delta_1 = 0$ and $\Delta_2 = 1/2\sigma^2$. The adjusted profile log-likelihood derivative is then

$$\frac{\partial l_\mathrm{p}}{\partial\sigma^2} + \frac{1}{2\sigma^2} = -\frac{n-1}{2\sigma^2} - \frac{(n-1)s^2}{2\sigma^4},$$

which gives the same adjusted log-likelihood as before.

5.2. *Example 3: Normal Mean, Variance Unknown*

The profile log-likelihood in the problem of normal mean and variance unknown is $-(n/2)\log(s_\mu^2/n) - n/2$, where $s_\mu^2 = \Sigma_1^n(y_i - \mu)^2$. The conditional profile log-likelihood is

$$-\{(n-2)/2\}\,\log(s_\mu^2/n) - n/2 - \tfrac{1}{2}\log(n/2).$$

The corresponding likelihood ratio statistics are $n\log\{1 + n(\bar y - \mu)^2/s^2\}$ and $(n-2)\log\{1 + n(\bar y - \mu)^2/s^2\}$, both monotone functions of the usual t-statistic. It is easily checked that the conditional profile log-likelihood ratio statistic has the correct scaling in the sense that its mean equals $1 + O(n^{-2})$ as opposed to $1 + O(n^{-1})$ for the profile log-likelihood ratio statistic.

Now $E\,U(\mu) = 0$,

$$\mathrm{var}(U(\mu)) = \frac{n^2}{2\sigma^2}\,\frac{\Gamma((n-2)/2)}{\Gamma(n/2)} \approx \frac{n^2}{\sigma^2}\,\frac{1}{n-2},$$

and $E\,\partial^2 l_\mathrm{p}(\mu)/\partial\mu^2 = n/\sigma^2$. Hence the adjusted profile log-likelihood $l_\mathrm{ap}(\mu)$ approximately equals the conditional profile log-likelihood.

Remark 3. It is interesting to consider whether we could use some other log-likelihood (instead of the profile log-likelihood) as the basis for deriving an adjusted log-likelihood. We might want to do this in cases where the profile log-likelihood is difficult to compute. An obvious candidate to use in place of the profile log-likelihood is $l(\psi, \hat\lambda)$, i.e. to fix λ at its overall maximizing value. The simplest problem in which $\hat\lambda_\psi$ varies with ψ (so that $l_\mathrm{p}(\psi) \neq l(\psi, \hat\lambda)$) is probably the normal means problem of example 3. Carrying through the adjustment procedure with $l(\psi, \hat\lambda)$ in place of $l_\mathrm{p}(\psi)$ gives

$$l_\mathrm{ap}(\psi) = -\frac{n}{2}\log s^2 - \frac{ns_\mu^2}{2s^2},$$

not equal to the marginal log-likelihood given earlier. Perhaps some other approximate profile log-likelihood would produce better results, but we have not investigated this further.

5.3. *Example 4: Weighted Normal Mean*

Suppose that we have q independent normal samples, with variance σ_j^2 and n_j observations in the jth sample. The parameter of interest is the common mean μ. Let

$\bar{y}_j$ and S_j be the sample mean and residual sum of squares from the jth sample. The profile log-likelihood for μ is

$$l_{\mathrm{p}}(\mu) = \sum_1^q -\frac{n_j}{2} \log\{S_j + n_j(\bar{y}_j - \mu)^2\}.$$

The conditional profile log-likelihood (given in Cox and Reid (1987)) replaces the first n_j in this expression by $n_j - 2$. The score function from the profile log-likelihood is

$$U(\mu) = \sum_1^q n_j^2 \frac{\bar{y}_j - \mu}{S_j + n_j(\bar{y}_j - \mu)^2}.$$

The mean of $U(\mu)$ is zero; this implies that the estimate derived from $l_{\mathrm{ap}}(\mu)$ will be the same as the maximum likelihood estimate $\hat{\mu}$, which is not true in general for the maximizer of l_{cp} in this problem. The variance correction factor can be shown to be approximately equal to

$$\sum \frac{n_j}{\sigma_j^2} \bigg/ \sum \frac{n_j^2}{\sigma_j^2(n_j - 2)}.$$

Thus, when the n_js and the σ_j^2s are equal, the adjusted profile log-likelihood approximately equals the conditional profile likelihood. Otherwise they are different.

We consider an example with $q = 4$, sample sizes 3, 4, 7 and 15, and standard deviations 1, 2, 5 and 8. Fig. 1 shows a typical result from a data set generated from this

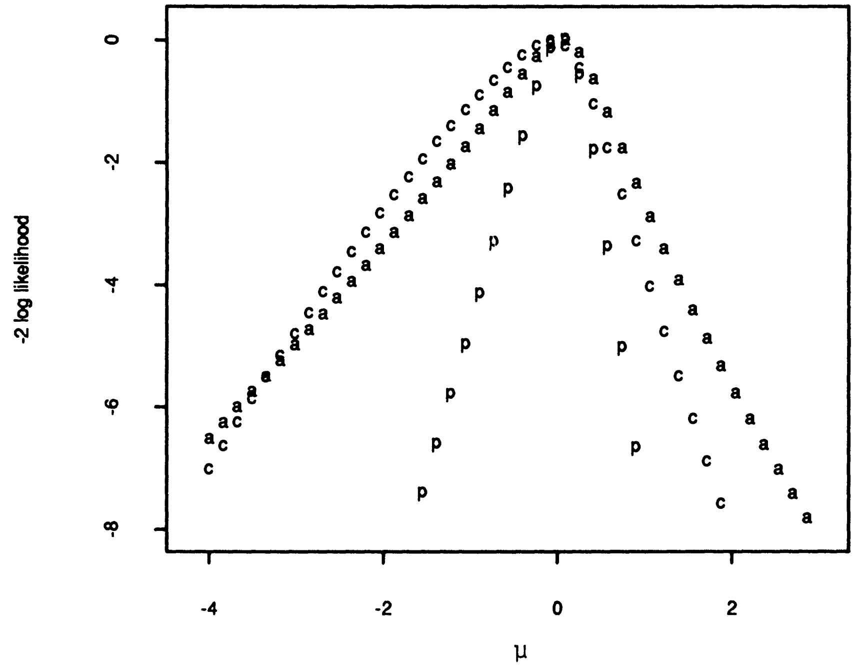

Fig. 1. 2{log-likelihood − max(log-likelihood)} for log-likelihoods in the weighted normal means problem: p, profile log-likelihood; c, conditional profile log-likelihood; a, adjusted profile log-likelihood

model. The 'p' represents the profile log-likelihood l_p, the 'c' the conditional profile log-likelihood l_{cp} and the 'a' the adjusted profile log-likelihood l_{ap}. Both l_{cp} and l_{ap} are wider than l_p, to account for the estimation of the variances. To investigate further, we ran a small simulation. For each of 1000 samples we computed the quantities $|\tilde{w}(0)-w(0)|$, $w(0)$ denoting the true log-likelihood ratio statistic obtained by setting the variances equal to their true values and $\tilde{w}(0)$ the log-likelihood statistic from either l_p, l_{cp} or l_{ap}. The quartiles of this quantity for l_p were (0.12, 0.45, 1.53), for l_{cp} (0.10, 0.33, 0.90) and for l_{ap} (0.09, 0.32, 0.87). The Monte Carlo standard error was about 0.03. We see that both the conditional and the adjusted profile log-likelihoods produce a log-likelihood ratio statistic that is considerably closer to the true log-likelihood ratio statistic than is that from the profile log-likelihood. Fig. 2 displays the various log-likelihood ratio statistics plotted against the true log-likelihood ratio statistic, for the first 200 simulations. The profile log-likelihood ratio statistic is often too large, a problem alleviated to some extent by the adjusted and conditional profile log-likelihoods.

5.4. *Example 5: Gamma Distribution*

The parameter of interest is the shape parameter ψ; calculations are most conveniently carried out by taking $\lambda = EY$ as the complementary parameter, λ being orthogonal to ψ (Cox and Reid (1987), section 3.2). Then the density of Y is

$$\left(\frac{\lambda}{\psi}\right)^{-\psi} \frac{y^{\psi-1}\exp(-\psi y/\lambda)}{\Gamma(\psi)}.$$

$\hat{\lambda}_\psi = \bar{y}$ for all ψ. The profile log-likelihood is

$$l_p(\psi) = n\psi \log \psi - n\psi \log \bar{y} + (\psi - 1) \sum_1^n \log y_i - n\psi - n \log \Gamma(\psi).$$

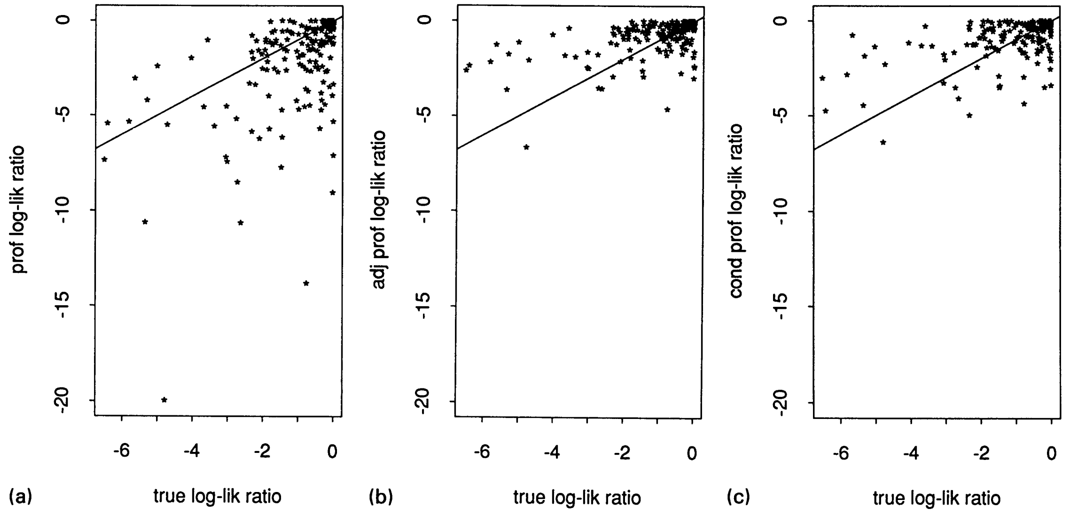

Fig. 2. Plots of log-likelihood ratios derived from (a) l_p, (b) l_{ap} and (c) l_{cp} versus the true log-likelihood ratio, for the many normal means problem: ———, 45° line

The conditional profile log-likelihood is

$$l_{\mathrm{cp}}(\psi) = l_{\mathrm{p}}(\psi) - \tfrac{1}{2} \log \psi.$$

If we use the standard approximation $\Gamma'(k)/\Gamma(k) \approx \log k - 1/2k$, then it can be shown that $m(\psi) \approx 1/2\psi$, $w(\psi) \approx 1$ and thus $l_{\mathrm{ap}}(\psi) \approx l_{\mathrm{cp}}(\psi)$. The term $\frac{1}{2} \log \psi$ can be viewed as a 'one degree of freedom' adjustment, analogous to the normal variance problem (Cox and Reid (1987), section 4.2.3; McCullagh and Nelder (1983)).

5.5. *Example 6: Binary Matched Pairs*

Assume that we have n pairs (Y_{i1}, Y_{i2}), with Y_{ij} equal to zero or unity, and the success probabilities π_{1i}, π_{i2} satisfy

$$\begin{aligned} \operatorname{logit} \pi_{1i} &= \lambda_i, \\ \operatorname{logit} \pi_{2i} &= \lambda_i + \psi. \end{aligned}$$

The log-odds ratio ψ is the parameter of interest and the λ_is are nuisance parameters. Let a, b, c and d denote the number of pairs of the form (0, 0), (0, 1), (1, 0) and (1, 1) respectively, with $a + b + c + d = n$. The standard conditional log-likelihood for this problem is formed by conditioning on $Y_{i1} + Y_{i2}$ and is given by

$$l_{\mathrm{c}}(\psi) = c\psi - (b + c) \log(1 + \exp \psi).$$

Only pairs of the form (1, 0) and (0, 1) enter the conditional log-likelihood. The profile log-likelihood is

$$l_{\mathrm{p}}(\psi) = \tfrac{1}{2}(c - b)\psi - (b + c) \log[\{1 + \exp(\psi/2)\}\{1 + \exp(-\psi/2)\}].$$

Barndorff-Nielsen (1986) gives a formula for the modified profile log-likelihood when $a = b = 0$:

$$l_{\mathrm{mp}}(\psi) = \tfrac{1}{2}(c - b)\psi - 3(b + c) \log\{\exp(\psi/4) + \exp(-\psi/4)\}.$$

The adjustment factors for the computation of $l_{\mathrm{ap}}(\psi)$ can be worked out analytically and involve cumbersome expressions. For small ψ the following expressions can be derived:

$$\left.\begin{aligned} \frac{\partial l_{\mathrm{p}}(\psi)}{\partial \psi} &= \frac{d - c}{2}\psi - \frac{\psi}{8} + O(\psi^2); \\ \frac{\partial l_{\mathrm{ap}}(\psi)}{\partial \psi} &= \frac{b - c}{2}\psi - \frac{3\psi}{16} + O(\psi^2); \\ \frac{\partial l_{\mathrm{mp}}(\psi)}{\partial \psi} &= \frac{b - c}{2}\psi - \frac{3\psi}{16} + O(\psi^2); \\ \frac{\partial l_{\mathrm{c}}(\psi)}{\partial \psi} &= \frac{b - c}{2}\psi - \frac{\psi}{4} + O(\psi^2). \end{aligned}\right\} \tag{14}$$

Thus both $l_{\mathrm{ap}}(\psi)$ and $l_{\mathrm{mp}}(\psi)$ adjust the maximum likelihood estimate in the correct direction, but only half of the way.

The approximate bias correction is equal to

$$-\sum_1^n \frac{\pi_{1i}(1-\pi_{1i})\pi_{2i}(1-\pi_{2i})(\pi_{1i}-\pi_{2i})}{\{\pi_{1i}(1-\pi_{1i})+\pi_{2i}(1-\pi_{2i})\}^2}$$

and when this is evaluated at $\hat{\lambda}_\psi$ we obtain

$$-\frac{1}{4}(b+c)\left\{\frac{\exp(-\psi/2)}{1+\exp(-\psi/2)}-\frac{\exp(\psi/2)}{1+\exp(\psi/2)}\right\}.$$

For small ψ this gives the same expression as $\partial l_{\text{ap}}(\psi)/\partial\psi$ in equations (14).

To compare the log-likelihoods we consider an example. Suppose that $a = 0$, $b = 0$, $c = 13$ and $d = 7$. Then the maximizing values of $l_c(\psi)$, $l_p(\psi)$, $l_{\text{mp}}(\psi)$ and $l_{\text{ap}}(\psi)$ are 0.62, 1.22, 0.82 and 0.83 respectively. Fig. 3 displays the various log-likelihoods.

5.6. *Example 7: Exponential Regression*

Suppose that we observe $Y_1, Y_2, \ldots Y_n$ from the exponential distribution with mean $\lambda\exp(-\psi z_i)$, the regression slope ψ being of interest. Following Cox and Reid (1987), section 4.2.2, we assume that $\Sigma z_i = 0$ which implies that ψ and λ are orthogonal. The profile log-likelihood is

$$l_p(\psi) = -n\log\sum_1^n y_i\exp(\psi z_i).$$

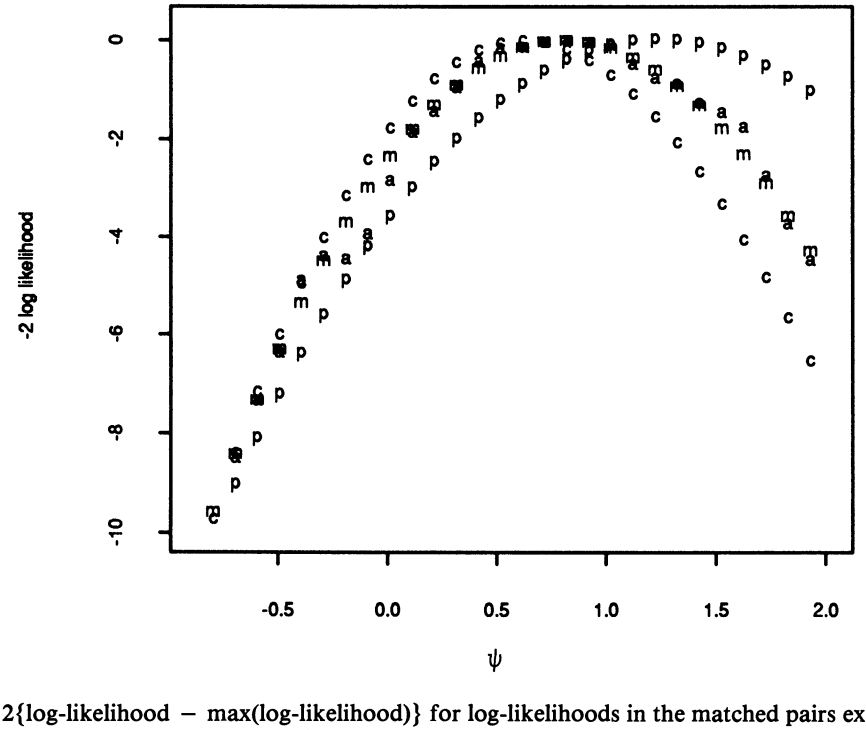

Fig. 3. 2{log-likelihood − max(log-likelihood)} for log-likelihoods in the matched pairs example: p, profile log-likelihood; c, conditional log-likelihood; m, modified profile log-likelihood; a, adjusted profile log-likelihood; the modified and adjusted log-likelihoods are very close and are much closer to the correct conditional log-likelihood than is the profile log-likelihood

The conditional profile log-likelihood replaces n by $n - 1$ while the modified profile likelihood replaces n by $n - 2$. As a result, all three log-likelihoods produce the same estimate but $l_{\mathrm{cp}}(\psi)$ and $l_{\mathrm{mp}}(\psi)$ make one and two degree of freedom adjustments (respectively) to the estimate of its precision. A simple calculation shows that $EU(\psi) \approx 0, E\{U(\psi)\}^2 \approx \Sigma_1^n z_i^2$,

$$E(\partial^2/\partial\psi^2)l_{\mathrm{p}}(\psi) = -\{n/(n+1)\} \sum_1^n z_i^2,$$

all expectations taken with respect to parameters $(\psi, \hat{\lambda}_\psi)$. Thus $l_{\mathrm{ap}}(\psi) \approx \{n/(n+1)\}\, l_{\mathrm{p}}(\psi)$, and we find that $l_{\mathrm{ap}}(\psi)$ adjusts $l_{\mathrm{p}}(\psi)$ by a lesser amount than do $l_{\mathrm{cp}}(\psi)$ and $l_{\mathrm{mp}}(\psi)$. However, the factors differ by $O(n^{-2})$ while the factors themselves are of size $O(n^{-1})$.

5.7. *Example 8: Normal Covariance Function Estimation*

Suppose that $Y \sim N_n\{X\beta, \Sigma(\psi)\}$ where ψ is the parameter of interest and β is a nuisance parameter. The profile log-likelihood for ψ is

$$l_{\mathrm{p}}(\psi) = -\tfrac{1}{2}\log(\det \Sigma) - \tfrac{1}{2} Q_2(R)$$

where

$$Q_2(R) = Y^{\mathrm{T}}\Sigma^{-1}Y - Y^{\mathrm{T}}\Sigma^{-1}X(X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}Y$$

is an invariant function of the residual vector

$$R = \{I - X(X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}\}\, Y.$$

The marginal log-likelihood based on any set of contrasts or residuals R is

$$l_{\mathrm{M}}(\psi) = -\tfrac{1}{2}\log(\det \Sigma) - \tfrac{1}{2}\log\{\det(X^{\mathrm{T}}\Sigma^{-1}X)\} - \tfrac{1}{2} Q_2(R),$$

which differs from the profile log-likelihood through the bias correction term

$$-\tfrac{1}{2}\log\{\det(X^{\mathrm{T}}\Sigma^{-1}X)\}.$$

The derivatives of the full log-likelihood are

$$\frac{\partial l}{\partial \beta} = X^{\mathrm{T}}\Sigma^{-1}(Y - X\beta)$$

$$\frac{\partial l}{\partial \psi^r} = -\tfrac{1}{2}\{\Sigma_{ij} - (Y^k - X_s^k\beta^s)(Y^l - X_s^l\beta^s)\Sigma_{ik}\Sigma_{jk}\}D_r^{ij},$$

where Σ^{ij} are the components of Σ, Σ_{ij} are the components of the inverse and $D_r^{ij} = \partial\Sigma^{ij}/\partial\psi^r$.

The exact corrections are difficult to derive but the first-order expressions are fairly simple to derive. For each fixed ψ, this is a full exponential family model. Hence application of equation (10) gives

$$E\left(\frac{\partial l_{\mathrm{p}}}{\partial \psi^r}\right) \simeq -D_r^{ij}\{\Sigma^{-1}X(X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}\}_{ij}$$

$$= \frac{1}{2}\frac{\partial}{\partial\psi^r}\log\{\det(X^{\mathrm{T}}\Sigma^{-1}X)\}.$$

Thus the first-order bias-corrected profile log-likelihood is

$$l_p(\psi) - \tfrac{1}{2} \log\{\det(X^T \Sigma^{-1} X)\},$$

which is identical with the marginal log-likelihood of the residual vector R, also called the restricted log-likelihood for ψ (Harville, 1974, 1977).

5.8. *Example 9: Comparison of Two Binomial Probabilities*

Suppose that $Y_1 \sim B(m_1, \pi_1)$, $Y_2 \sim B(m_2, \pi_2)$ are independent random variables with

$$\operatorname{logit} \pi_1 = \lambda + \psi,$$
$$\operatorname{logit} \pi_2 = \lambda.$$

The full log-likelihood is

$$y_1(\lambda + \psi) + y_2\lambda - m_1 \log\{1 + \exp(\lambda + \psi)\} - m_2 \log(1 + \exp \lambda)$$

and the derivatives are

$$\frac{\partial l}{\partial \psi} = y_1 - m_1 \exp(\lambda + \psi)/\{1 + \exp(\lambda + \psi)\} = y_1 - \mu_1$$

$$\frac{\partial l}{\partial \lambda} = y_1 - \mu_1 + y_2 - \mu_2.$$

The maximum likelihood estimate $\hat{\lambda}_\psi$ satisfies

$$\frac{\hat{\mu}_1(m_2 - \hat{\mu}_2)}{\hat{\mu}_2(m_1 - \hat{\mu}_1)} = \exp \psi,$$

where

$$\hat{\mu}_1 \equiv \hat{\mu}_1(\psi) = m_1 \exp(\hat{\lambda}_\psi + \psi)/\{1 + \exp(\hat{\lambda}_\psi + \psi)\},$$

and similarly for $\hat{\mu}_2(\psi)$.

From the calculations in Section 4 we find that the first-order expectation of $\partial l/\partial\psi$ evaluated at $\hat{\lambda}_\psi$ is

$$E\left(\frac{\partial l_{\text{pl}}}{\partial \psi}\right) \simeq -\frac{\sigma_1^2 \sigma_2^2}{(\sigma_1^2 + \sigma_2^2)^2} (\pi_2 - \pi_1)$$

when $\sigma_1^2 = m_1 \pi_1 (1 - \pi_1)$ and similarly for σ_2^2. Thus the bias-corrected log-likelihood derivative is

$$y_1 - \hat{\mu}_1(\psi) + \frac{\hat{\sigma}_1^2 \hat{\sigma}_2^2}{(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)^2} (\hat{\pi}_2 - \hat{\pi}_1) \qquad (15)$$

with $\hat{\sigma}_1^2 = m_1 \hat{\pi}_1 (1 - \hat{\pi}_1)$. We now show that expression (15) is approximately equal to the derivative of the conditional log-likelihood for $Y_1 \mid Y_.$.

The derivative of the conditional log-likelihood is

$$\frac{\partial l_c}{\partial \psi} = y_1 - \mu_c(\psi)$$

where $\mu_c(\psi)$, the conditional mean of Y_1 given $Y_.$, satisfies

$$\frac{\mu_c(\psi)\{m_2 - y_. + \mu_c(\psi)\} + \kappa_2(\psi)}{\{y_. - \mu_c(\psi)\}\{m_1 - \mu_c(\psi)\} + \kappa_2(\psi)} = \psi$$

with $\kappa_2(\psi) = \mathrm{var}(Y|\, Y_.)$. Thus, for small ψ and for large m_1 and m_2,

$$\mu_c(\psi) \simeq \hat{\mu}_1 + \kappa_2\psi/(m_1 + m_2) + O(\psi^2)$$

$$\kappa_2 \simeq \sigma_1^2\sigma_2^2/(\sigma_1^2 + \sigma_2^2)$$

and

$$\mu_c(\psi) - \hat{\mu}_1 \simeq \frac{\sigma_1^2\sigma_2^2}{(\sigma_1^2 + \sigma_2^2)^2}(\pi_1 - \pi_2) \simeq -E\left(\frac{\partial l_p}{\partial\psi}\right).$$

In other words, the first-order bias correction approximates the conditional log-likelihood derivative.

5.9. *Example 10: Ratio of Normal Means*

For the ratio of normal means we observe that $Y_1 \sim \mathcal{N}(\eta_1, 1/n)$, $Y_2 \sim \mathcal{N}(\eta_2, 1/n)$ and we would like to make inferences about $\psi = \eta_2/\eta_1$. The radius $\lambda = \sqrt{(\eta_1^2 + \eta_2^2)}$ is orthogonal to ψ. The profile log-likelihood is

$$l_p(\psi) = -\frac{n}{2(1 + \psi^2)}(y_1\psi - y_2)^2.$$

We could argue that this is a reasonable log-likelihood because the corresponding log-likelihood ratio statistic is the square of the Euclidean distance between (y_1, y_2) and the level curve $C_\psi = \{(\eta_1, \eta_2);\ \eta_2/\eta_1 = \psi\}$. This distance is easily seen to be distributed $\mathcal{N}(0, 1)$ and thus the log-likelihood ratio statistic is exactly χ_1^2. The likelihood intervals also agree with those derived by Fieller (1954) and Creasy (1954).

We can easily check that $l_{cp}(\psi) = l_p(\psi)$, i.e. that the correction term is zero. However, $\lambda = \log\sqrt{(\eta_1^2 + \eta_2^2)}$ (say) is also orthogonal to ψ and results in a conditional profile log-likelihood that is different from $l_p(\psi)$, by a term of magnitude $O_p(n^{-1})$. A fairly cumbersome calculation shows that $E\{\partial l_p(\psi)/\partial\psi\} = 0$ and the variance correction is $w(\psi) = 1 + O(n^{-1})$. However, the likelihood ratio statistic from $l_{ap}(\psi)$ can be shown to differ from that of $l_p(\psi)$ by only $O_p(n^{-1})$. Kalbfleisch and Sprott (1970) discuss some other likelihoods for this problem.

6. JUSTIFICATION OF ADJUSTED PROFILE LIKELIHOOD

One justification for the adjusted profile likelihood can be found in the theory of optimal estimating equations. In particular, unbiasedness of the estimating equation essentially guarantees consistency, and the condition on the derivative matrix guarantees asymptotic optimality within a limited class of estimating functions. For further details we refer the reader to Godambe (1960), Godambe and Thompson (1974), Lindsay (1982) and Godambe and Heyde (1987).

A more obvious heuristic justification for this procedure is that the resulting score statistic for ψ would have approximately the appropriate mean and variance, which is important for establishing large sample results for maximum likelihood estimates and the likelihood ratio test. Hence we argue that the centring of the profile log-likelihood score function should improve the consistency of the maximizer of the likelihood, while the rescaling should improve the second-derivative approximation to its variance and the chi-square approximation to the distribution of the log-likelihood ratio statistic. We have no strong argument for this claim, but only the following heuristics.

In more detail, the construction of the adjusted profile log-likelihood is based on the requirements

$$E_{\psi, \hat{\lambda}_\psi} \tilde{U} = 0, \tag{16}$$

$$\operatorname{var}_{\psi, \hat{\lambda}_\psi}(\tilde{U}(\psi)) = -E_{\psi, \hat{\lambda}_\psi} \frac{\partial}{\partial \psi} \tilde{U}(\psi) \tag{17}$$

for all ψ, where $\tilde{U}(\psi) = \{U(\psi) - m(\psi)\}/w(\psi)$. To study further whether these requirements are reasonable, it is useful to consider two separate cases: problems for which the maximum likelihood estimate $\hat{\psi}$ is inconsistent and problems for which it is consistent. Throughout let (ψ_0, λ_0) denote the true values of the parameters.

In inconsistent cases, $E_{\psi_0, \lambda_0}(\partial/\partial\psi) l_p(\psi) \neq 0$ and requirement (16) attempts to correct it. However, requirement (16) ensures that the expectation of the score with respect to parameters $(\psi, \hat{\lambda}_\psi)$, rather than (ψ_0, λ_0), is zero. Thus for $\tilde{\psi}$ (the maximizer of $l_{ap}(\psi)$) to be consistent, we need either $\hat{\lambda}_{\psi_0} \rightarrow \lambda_0$ or that the moments in requirements (16) and (17) do not depend on λ. If either of these holds, requirement (17) ensures that the log-likelihood is calibrated correctly in the sense described in Section 3. Typically, however, the first of these conditions will not be satisfied, since in cases when $\hat{\psi}$ is inconsistent the dimension of λ tends to infinity. An example is the many normal means problems in which the estimates of the pairwise means are not consistent. However, the second condition will be satisfied when $l_{ap}(\psi)$ corresponds to a marginal log-likelihood for ψ. This is the case in the many normal means problem. Beyond this, it seems difficult to justify $l_{ap}(\psi)$ in inconsistent cases. For example the consistency of $\tilde{\psi}$ in the matched pairs example is an open question.

In cases for which $\hat{\psi}$ is consistent it is of interest to compare the adjusted profile log-likelihood with the conditional profile likelihood of Cox and Reid. We are able to make this comparison only in the special case of exponential family canonical parameters. In Section 4 we showed that in that case the bias correction for the score is $b_r = -\frac{1}{2}\kappa_{r,i,j}\kappa^{i,j}$. Now the profile log-likelihood correction given by Cox and Reid is

$$\tfrac{1}{2} \log\{\det n j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)\} \tag{18}$$

where $j_{\lambda\lambda}$ is the observed Fisher information per observation for the λ components. It is easy to show that the derivative of expression (18) is b_r. Hence examination of equation (13) reveals that the adjusted profile log-likelihood score differs from the conditional profile likelihood score by the additional $O(1)$ term

$$\frac{\Delta_1}{\kappa_{r,s} - \Delta_2} \frac{\partial l_p}{\partial \psi^r}.$$

This expression is zero in the normal and Poisson cases, but not in general.

It is interesting that the adjusted profile log-likelihood, like the conditional profile log-likelihood, seems automatically to make the Bartlett adjustments in the Gaussian problems considered earlier. We have not investigated whether Bartlett adjustments might improve the χ_1^2 approximation to the distribution of the log-likelihood ratio statistic from $l_{\mathrm{ap}}(\psi)$.

ACKNOWLEDGEMENTS

We would like to thank David Andrews, John Chambers, Bradley Efron, Rob Kass, Richard Lockhart and Nancy Reid for helpful discussions, and the associate editor and referee for valuable comments. The second author was supported by the Natural Sciences and Engineering Research Council of Canada.

APPENDIX A: DERIVATION OF ASYMPTOTIC APPROXIMATION FOR MEAN ADJUSTMENT

The Taylor expansion for the profile log-likelihood about the true parameter point is

$$\begin{aligned} l_{\mathrm{p}}(\psi) &= l(\psi,\, \hat{\lambda}_\psi) = \sup_\lambda l(\psi,\, \lambda) \\ &= l(\psi,\, \lambda) + \frac{\partial l}{\partial \lambda}(\hat{\lambda}_\psi - \lambda) + \frac{1}{2}\frac{\partial^2 l}{\partial \lambda^2}(\hat{\lambda}_\psi - \lambda)^2 + \ldots . \end{aligned}$$

Differentiation with respect to ψ gives

$$\begin{aligned} \frac{\partial l_{\mathrm{p}}}{\partial \psi} &= \frac{\partial}{\partial \psi} l(\psi,\, \hat{\lambda}_\psi;\, y) \\ &= \frac{\partial l}{\partial \psi} + \frac{\partial^2 l}{\partial \psi\, \partial \lambda}(\hat{\lambda}_\psi - \lambda) + \frac{1}{2}\frac{\partial^3 l}{\partial \psi\, \partial \lambda^2}(\hat{\lambda}_\psi - \lambda)^2 + \ldots + \frac{\partial}{\partial \lambda} l(\psi,\, \hat{\lambda}_\psi;\, y)\frac{\partial \hat{\lambda}_\psi}{\partial \psi}. \end{aligned}$$

Under the usual regularity conditions for large n, the first three terms are $O_{\mathrm{p}}(n^{1/2})$, $O_{\mathrm{p}}(n^{1/2})$ and $O_{\mathrm{p}}(1)$ respectively. The first term has zero mean but the remaining two have mean $O(1)$ if $\hat{\lambda}_\psi$ is a consistent estimate of λ.

Using index notation with r denoting components of ψ and i, j, k, ... denoting components of λ and $\hat{\lambda}_\psi$, we have

$$\frac{\partial l_{\mathrm{p}}}{\partial \psi^r} = U_r + U_{ri}(\hat{\lambda}^i - \lambda^i) + \tfrac{1}{2} U_{rij}(\hat{\lambda}^i - \lambda^i)(\hat{\lambda}^j - \lambda^j) + \ldots .$$

$\hat{\lambda}^i$ are the components of $\hat{\lambda}_\psi$, not the components of the overall maximum likelihood estimate. The expansion for $\hat{\lambda}^j - \lambda^j$ in terms of log-likelihood derivatives is

$$\hat{\lambda}^i - \lambda^i = \kappa^{i,j} U_j + \kappa^{i,j}\kappa^{k,l}(U_{jk} - \kappa_{jk}) U_l + \tfrac{1}{2}\kappa^{ijk} U_j U_k + \ldots$$

where $\kappa_{i,j}$ are the components of the Fisher information for λ with ψ fixed and $\kappa^{i,j}$ are the components of the inverse matrix. On substituting these into the expansion for $\partial l_{\mathrm{p}}/\partial \psi^r$ we find

$$\frac{\partial l_{\mathrm{p}}}{\partial \psi^r} = U_r + U_{ri}\{\kappa^{i,j} U_j + \kappa^{i,j}\kappa^{k,l}(U_{jk} - \kappa_{jk}) U_l + \tfrac{1}{2}\kappa^{ijk} U_j U_k\} + \tfrac{1}{2} U_{rij}\kappa^{i,k}\kappa^{j,l} U_k U_l + O_{\mathrm{p}}(n^{-1/2}).$$

The expectation of this expression is

$$\begin{aligned} E\left(\frac{\partial l_{\mathrm{p}}}{\partial \psi^r}\right) &= \kappa_{ri,j}\kappa^{i,j} + \kappa_{ri}\kappa^{i,j}\kappa^{k,l}\kappa_{jk,l} + \tfrac{1}{2}\kappa_{ri}\kappa^{ijk}\kappa_{j,k} + \tfrac{1}{2}\kappa_{rij}\kappa^{i,j} + O(n^{-1/2}) \\ &= (\kappa_{ri,j} - \kappa_{r,k}\kappa^{k,l}\kappa_{li,j})\kappa^{i,j} + \tfrac{1}{2}(\kappa_{rij} - \kappa_{r,k}\kappa^{k,l}\kappa_{lij})\kappa^{i,j}. \end{aligned}$$

On using the third-order Bartlett identity

$$\kappa_{rij} = -\kappa_{r,i,j} - \kappa_{r,ij} - \kappa_{ri,j} - \kappa_{rj,i},$$

we obtain finally

$$E\left(\frac{\partial l_{\mathrm{p}}}{\partial \psi^r}\right) \simeq -\tfrac{1}{2}(\kappa_{r,i,j} - \kappa_{r,k}\kappa^{k,l}\kappa_{l,i,j})\kappa^{i,j} - \tfrac{1}{2}(\kappa_{r,ij} - \kappa_{r,k}\kappa^{k,l}\kappa_{l,ij})\kappa^{i,j}.$$

REFERENCES

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

——— (1986) Inference on full or partial parameters, based on the standardized log likelihood ratio. *Biometrika*, **73**, 307-322.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*. Belmont: Wadsworth.

Cox, D. R. and Hinkley, D. (1974) *Theoretical Statistics*, ch. 2. London: Chapman and Hall.

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

Creasy, M. A. (1954) Limits for the ratio of means. *J. R. Statist. Soc.* B, **16**, 186-194.

Fieller, E. C. (1954) Some problems in interval estimation. *J. R. Statist. Soc.* B, **16**, 175-185.

Godambe, V. P. (1960) An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.*, **31**, 1208-1211.

Godambe, V. P. and Heyde, C. C. (1987) Quasi-likelihood and optimal estimation. *Int. Statist. Rev.*, **55**, 231-244.

Godambe, V. P. and Thompson, M. E. (1974) Estimating equations in the presence of a nuisance parameter. *Ann. Statist.*, **2**, 568-571.

Harville, J. A. (1974) Bayesian inference for variance components using error contrasts. *Biometrika*, **61**, 383-385.

——— (1977) Maximum likelihood approaches to variance component estimation and to related problems (with discussion). *J. Am. Statist. Ass.*, **72**, 320-340.

Kalbfleisch, J. D. and Sprott, D. A. (1970) Application of likelihood methods to models involving large numbers of parameters (with discussion). *J. R. Statist. Soc.* B, **32**, 175-208.

Lindsay, B. (1982) Conditional score functions: some optimality results. *Biometrika*, **69**, 503-512.

McCullagh, P. (1987) *Tensor Methods in Statistics*. London: Chapman and Hall.

McCullagh, P. and Nelder, J. A. (1983) *Generalized Linear Models*, p. 157. London: Chapman and Hall.

Silvey, S. D. (1975) *Statistical Inference*, p. 36. Penguin: Harmondsworth.

Sweeting, T. J. (1987) Discussion on Parameter orthogonality and approximate conditional inference (by D. R. Cox and N. Reid). *J. R. Statist. Soc.* B, **49**, 20-21.