Aspects of modified profile likelihood

N. Reid

Department of Statistics, University of Toronto Toronto, Canada M5S 1A1

Abstract

A modification of profile likelihood was suggested in Cox and Reid (1987), which is similar to but somewhat simpler than that of Barndorff-Nielsen (1983). The modification attempts to adjust for the presence of nuisance parameters, and to approximate the appropriate marginal or conditional likelihood when such is available. Some recent results on the modified profile likelihood are outlined, particularly emphasizing computation and asymptotic properties. Some examples are used to illustrate inference using modified profile likelihoods.

KEYWORDS: approximate likelihood, asymptotic inference, conditional inference, likelihood, profile likelihood, modified profile likelihood

1. INTRODUCTION

In this paper we assume that our response vector \( Y \) follows a parametric model with density or probability function \( f_Y(y; \psi, \lambda) \), where \( \psi \) is a scalar parameter of interest and \( \lambda \) is a vector of nuisance parameters. Examples include several \( 2 \times 2 \) tables with a common odds ratio \( \psi \), and different marginal probabilities in each table; several normally distributed samples with common mean \( \psi \) but different variances in each sample; and linear regression models with interest focussing on a particular contrast or regression coefficient, the remaining regression coefficients and unknown variance being treated as nuisance parameters.

The log-likelihood function, which will also be called just the likelihood function is given by

\[
I(\psi, \lambda) = I(\psi, \lambda; y) = \log f_Y(y; \psi, \lambda) + c
\]

and is typically used to evaluate the plausibility of a range of values of \((\psi, \lambda)\) given an observed data vector \( y \).
In the special situation when there are in fact no unknown nuisance parameters, and the likelihood function depends only on $\psi$ and $y$, inference for $\psi$ is particularly straightforward. For example $l(\psi) = l(\psi; y)$ is easily plotted as a function of $\psi$, and the plausibility of different values can be assessed visually. Various summary statistics are available, the most usual of which are the maximum likelihood estimate $\hat{\psi}$, the score function $U(\psi)$, the likelihood ratio statistic $w(\psi)$ and its signed square root $r(\psi)$:

\[
\hat{\psi} = \arg \sup \ l(\psi; y) \\
U(\psi) = \frac{\partial}{\partial \psi} l(\psi; y) \\
w(\psi) = 2\{l(\hat{\psi}) - l(\psi)\} \\
r(\psi) = \text{sign}(\hat{\psi} - \psi)\{w(\psi)\}^{1/2}.
\]

Approximate confidence intervals can be constructed using any of the first order asymptotic results

\[
(\hat{\psi} - \psi)\{-l''(\hat{\psi})\}^{1/2} \xrightarrow{d} N(0, 1) \\
U(\psi)\{-l''(\psi)\}^{-1/2} \xrightarrow{d} N(0, 1) \\
w(\psi) \xrightarrow{d} \chi^2_1 \\
r(\psi) \xrightarrow{d} N(0, 1)
\]

which are valid in i.i.d. sampling as the dimension of $Y$ goes to $\infty$, and in more general models as the expected Fisher information $i(\psi) = E\{-l''(\psi); \psi\}$ goes to $\infty$. In addition, much more accurate confidence intervals can be constructed from the approximation

\[
\text{pr}\{R(\psi) \leq r\} = \Phi(r) + \frac{1}{q} \left( \frac{1}{r} - \frac{1}{q} \right)
\]

where $r = r(\psi)$ and $q = q(\psi) = (\hat{\psi} - \psi)\{-l''(\hat{\psi})\}^{1/2}$. This approximation, often called the Lugannani and Rice approximation, is a third order asymptotic result, as the error is $O(n^{-3/2})$ for values of $\psi$ within $1/\sqrt{n}$ neighbourhoods of $\hat{\psi}$, whereas the error in (2) is $O(n^{-1/2})$. The version given in (3) is appropriate for exponential families. Other versions are reviewed in Reid and Fraser (1991). An alternative third order asymptotic result can be obtained by Bartlett correction of $w(\psi)$. Writing $\tilde{w}(\psi) = w(\psi)/E\{w(\psi); \psi\}$, it can be shown that $\tilde{w}(\psi)$ is approximately distributed as $\chi^2_1$, with error $O(n^{-2})$ (Cox and Hinkley, 1974, p.339; Barndorff-Nielsen and Hall, 1988).

The goal in this paper is to discuss procedures for constructing a likelihood function or approximate likelihood function for the parameter of interest $\psi$, in the presence of several unknown nuisance parameters $\lambda$. The advantages of such a function, if it can be obtained, are that the simple and accurate methods for inference about $\psi$ discussed
above may become available. We will concentrate mainly on the profile likelihood and the modification of it suggested by Cox and Reid (1987).

2. PROFILE LIKELIHOOD AND MODIFICATIONS

A simple and widely used method for constructing a function of $\psi$ only, from the likelihood function $l(\psi, \lambda)$ is to maximize over the nuisance parameter $\lambda$. Letting $\hat{\lambda}_\psi$ be the maximum likelihood estimate of $\lambda$ for a fixed value of $\psi$, the profile likelihood is defined by:

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

Although $l_p$ is not a likelihood function, in the sense that it is not the (log of) a density function for an observable random variable, it has some features of the one-parameter likelihood function discussed in Section 1. For example, if we define the summary statistics as in (1) but with $l$ replaced by $l_p$, we have that $\hat{\psi} = \tilde{\psi}_p$, and the first, third and fourth asymptotic results of (2) hold, as the sample size or Fisher information goes to $\infty$, assuming the dimension of $\lambda$ stays fixed. However, if the dimension of $\lambda$ increases with the sample size, then these asymptotic results are no longer valid in general. For example, $\hat{\psi}$ may be inconsistent or inefficient, $U_p(\psi)$ may not have exact or asymptotic mean 0, or the limiting distribution of $r_p(\psi)$ may not be $N(0,1)$.

**Example 1.** Suppose $(Y_{i1}, \ldots, Y_{im})$ are i.i.d. random variables from a $N(\lambda_i, \psi)$ distribution, for $i = 1, \ldots, n$. Then as $n \to \infty$, $\hat{\psi}$ is not consistent for $\psi$. In fact, if $m_i = 2$, $\hat{\psi}$ converges in probability to $\psi/2$. Alternatively, if $(Y_{i1}, \ldots, Y_{im})$ are $N(\psi, \lambda_i)$, and the parameter of interest is the mean, then $\hat{\psi}$ is consistent but not efficient. The maximum likelihood estimator $\hat{\psi}$ is defined as the solution of the equation $\sum m_i^2 (Y_i - \psi) / \{S_i^2 + (Y_i - \psi)^2\} = 0$, and it is easily verified that the asymptotic variance of the estimate is larger than that of the solution of $\sum m_i(m_i - 2)(Y_i - \psi) / \{S_i^2 + (Y_i - \psi)^2\} = 0$. Both these examples are due to Neyman and Scott (1948), and classes of problems with the dimension of the nuisance parameter increasing with the sample size are often called Neyman-Scott problems.

**Example 2.** Suppose that $(Y_{11}, Y_{12})$ follow Bernoulli distributions with parameters $p_{1i}$ and $p_{2i}$, $i = 1, \ldots, n$, and that $\psi = \log[(p_{1i}(1 - p_{2i})]/[p_{2i}(1 - p_{1i})]$ is the common odds ratio, with $\lambda_i = \log(p_{2i}/(1 - p_{2i}))$, say. Then again, as $n \to \infty$, $\hat{\psi}$ converges in probability to $\psi/2$, cf. for example, Breslow and Day (1980, p.250).

**Example 3.** We take $n$ pairs of observations, each with exponential distributions with mean $\psi \lambda_i$ and $\psi / \lambda_i$, respectively, so the parameter of interest is a product of means. It is not hard to verify that $\hat{\psi}$ converges in probability to $(\pi/4)\psi$ (Cox and Reid, 1991b).

These examples suggest the possibility that even if the dimension of $\lambda$ is fixed, as $n \to \infty$, the profile likelihood may not provide accurate inference for $\psi$. For example, the maximum likelihood estimate may have large bias, or the curvature of the profile likelihood may provide a misleading estimate of the precision of the maximum likelihood estimate, etc.
A relatively simple modification suggested in Cox and Reid (1987) is to define

\[ l_m(\psi) = l(\psi, \hat{\lambda}) - \frac{1}{2} \log |j_{\lambda\lambda}(\psi, \hat{\lambda})| \]

\[ = l_p(\psi) - c(\psi), \tag{4} \]

say, where \( j_{\lambda\lambda}(\psi, \lambda) = -\partial^2 l(\psi, \lambda)/\partial \lambda \partial \lambda^T \) is the nuisance parameter submatrix of the observed information matrix.

The function \( l_m \), although not a likelihood function in the sense mentioned above, can be used to compute the summary statistics (1), with \( l(\psi) \) there replaced by \( l_m(\psi) \). In examples that have been considered, the asymptotic results outlined in (2) seem to provide better approximations than those based on \( l_p(\psi) \); i.e., there is less bias in \( \psi_m \), the \( \chi^2_1 \) approximation to the distribution of \( w_m(\psi) \) seems more accurate, etc. The function \( l_m \) is an approximate likelihood function, in that it approximates the likelihood for a conditional density \( f(\psi|\hat{\lambda}_\psi) \) to \( O(n^{-1}) \), under the restriction that \( \lambda \) is orthogonal to \( \psi \), i.e., the expected Fisher information matrix is block diagonal. Under parameter orthogonality, \( l_m \) also approximates, to \( O(n^{-1}) \), the log of the ratio of the posterior density for \( \psi \) to the prior density for \( \psi \), if the priors for \( \psi \) and \( \lambda \) are independent.

A major disadvantage of \( l_m \) is that it is not invariant to reparametrizations of the nuisance parameter \( \lambda \). The lack of invariance is \( O(n^{-1}) \) if \( \psi \) and \( \lambda \) are orthogonal, and is thus smaller of order of magnitude than the terms retained in (4). As a result, (4) is only appropriate when \( \psi \) and \( \lambda \) are orthogonal. Although any given nuisance parameter vector can be redefined so that it is orthogonal to a scalar parameter of interest, it may be difficult to obtain the reexpression explicitly.

**Example 4.** Suppose \( Y_{ij} \) follows the following \( AR(1) \) model:

\[ Y_{ij} - \mu_i = \psi(Y_{i,j-1} - \mu_i) + \epsilon_{ij} \quad j = 1, \ldots, r; \quad i = 1, \ldots, m \]

with \( \epsilon_{ij} \) distributed as \( N(0, \sigma^2) \). The nuisance parameter vector \( (\mu, \sigma^2) \) can be reparametrized as \( (\mu, \lambda) \), where \( \lambda = \log\{\sigma(1 - \psi^2)^{1/(2r)}\} \). This example is considered in detail in Craddock, Reid and Cox (1989), where it is shown that \( \hat{\psi}_m \) has smaller bias than \( \hat{\psi}_p \), and that \( w_m \) is more nearly distributed as \( \chi^2_1 \) than \( w_p \). In addition, in this problem there exists a statistic whose marginal distribution depends only on \( \psi \), and \( l_m \) is in fact the log-likelihood for this marginal distribution. This latter point is also discussed in Tunnicliffe-Wilson (1989) and Bellhouse (1990). However, if the nuisance parameter is expressed as \( (\mu, \lambda^*) \), with \( \lambda^* = \sigma(1 - \psi^2)^{(1/2r)} \), the improvements in the application of the first order theory are much smaller. That is, \( l_m \) has better behaviour than \( l_p \), but not as good as \( l_m \) using \( \lambda \). Thus the lack of invariance to \( O(n^{-1}) \) can be important in applications.

The drawbacks associated with \( l_m \) are overcome by a different modification to profile likelihood due to Barndorff-Nielsen (1983). It is defined as

\[ l_{BN}(\psi) = l(\psi, \hat{\lambda}_\psi) - \frac{1}{2} \log |j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)| + a(\psi) \tag{5} \]
where \( a(\psi) \) is an adjustment factor that ensures that \( l_{BN} \) is invariant to reparametrizations of the nuisance parameter \( \lambda \). Thus parameter orthogonality is not required. The function \( a(\psi) \) is defined to be

\[
a(\psi) = \left| \frac{d\hat{\lambda}_\psi}{d\lambda} \right|
\]

where \( \hat{\lambda} \) is the overall maximum likelihood estimate of \( \lambda \). As indicated in Barndorff-Nielsen (1983, 1987), in addition to being invariant to the nuisance parametrization, \( l_{BN} \) approximates the conditional density of \( f(y|\hat{\lambda}_\psi) \) in models for which conditional inference is appropriate, and also can approximate the marginal density \( f(u_\psi) \) in models for which marginal inference is appropriate (i.e. in models for which there exists a statistic \( u_\psi \) whose marginal distribution depends only on \( \psi \)). This approximation is generally valid to \( O(n^{-1}) \), and in models with special structure, such as exponential family models, is accurate to \( O(n^{-3/2}) \).

A disadvantage of \( l_{BN} \) is that the correction factor \( a(\psi) \) is in general rather difficult to compute. In particular, it is necessary to be able to write \( \hat{\lambda}_\psi \) as a function of \( \hat{\lambda} \) and some complementary statistic, which may depend on \( \psi \). A further difficulty with \( l_{BN} \) is that the ‘densities’ \( f(y|\hat{\lambda}_\psi) \) and \( f(u_\psi) \) do not really provide likelihood functions, because there is a further dependence on \( \psi \) in the differential element. This point is further elaborated in Fraser and Reid (1989).

3. PROPERTIES OF \( l_m(\cdot) \)

In this section we consider properties of the likelihood summary statistics (1) based on \( l_m(\cdot) \), in order to provide some indication of when and why the modification indicated in (4) is effective.

The modified score statistic, \( U_m = l'_m(\psi) = l'_p(\psi) - c'(\psi) \) has the following stochastic expansion (Ferguson, Reid and Cox, 1991):

\[
l'_m(\psi) = \sqrt{n}a_{11} + a_{12} + \frac{1}{\sqrt{n}}a_{13}
- \frac{1}{2}(a_{20} + \frac{1}{\sqrt{n}}a_{21}) + O_p(n^{-1}),
\]

where the first three terms come from the expansion of \( l'_p(\psi) \) and the remaining terms come from the expansion of the correction term. The functions \( a_{ij} \) depend on observed and expected derivatives of the full log-likelihood function \( l(\psi, \lambda) \). For example, writing \( Z_{\psi\lambda} = (1/\sqrt{n})[\partial l/\partial \psi \lambda - E\{\partial l/\partial \psi \lambda\}] \), and \( nI_{\psi\lambda\lambda} \) for the expected third derivative, etc., the terms above are expressed as

\[
a_{11} = Z_{\psi} = \frac{1}{\sqrt{n}}\partial l(\psi, \lambda)/\partial \psi,
\]

\[
a_{12} = \frac{Z_{\psi\lambda}Z_{\lambda}}{I_{\lambda\lambda}} + \frac{1}{2} \frac{Z_{\lambda}^2}{I_{\lambda\lambda}^2},
\]

\[
a_{20} = \frac{I_{\psi\lambda\lambda}}{I_{\lambda\lambda}^2},
\]
and, in particular $Ea_{11} = Ea_{13} = Ea_{21} = 0$, so that

$$EU_p(\psi) = -\frac{1}{2} \frac{I_{\psi\psi\psi}}{I_{\psi\psi}} = O(1)$$

$$EU_m(\psi) = O(n^{-1}).$$

That is, the score function from $l_m$ is more nearly unbiased for $0$ than is that from the profile likelihood function. The same result was obtained by Liang (1987) by a different route. McCullagh and Tibshirani (1990) consider direct adjustments to $U_p$ to obtain this unbiasedness property. It is unfortunately not the case that $E - U_m(\psi) = \text{var}\{U_m(\psi)\}$, even to $O(n^{-1})$, a kind of information unbiasedness of score functions from genuine likelihoods. McCullagh and Tibshirani (1990) also consider adjustments to $U_p$ to attain this property as well.

Similar but slightly weaker results can be obtained for the maximum likelihood estimate. Recall that $\hat{\psi}$ is the solution of $U_p(\psi) = 0$, and we denote by $\hat{\psi}_m$ the solution of $U_m(\psi) = 0$. The stochastic expansion for $\hat{\psi}$ is

$$\sqrt{n}(\hat{\psi} - \psi) = Z_{\psi} \frac{b_{11}}{I_{\psi\psi}} + \frac{1}{\sqrt{n}}(b_{11} + b_{12}) + O_p(n^{-1})$$

where the $b_{11}$ terms involve only $\psi$-derivatives of $l(\psi, \lambda)$ and the $b_{12}$ terms involve derivatives with respect to $\psi$ and $\lambda$. The $b_{11}$ terms are the same as those arising in the stochastic expansion of $\hat{\psi}$ in the case of no nuisance parameters; the nuisance parameter component is entirely determined by $b_{12}$. The stochastic expansion for $\hat{\psi}_m$, given in Ferguson (1991), is

$$\sqrt{n}(\hat{\psi}_m - \psi) = Z_{\psi} \frac{b_{11} + b_{12} + b_{13}}{\sqrt{n}} + O_p(n^{-1}),$$

where

$$b_{13} = -\frac{1}{2} \frac{I_{\psi\psi\psi}}{I_{\psi\psi}} = -Eb_{12}$$

so that, in expectation, the second order term in the expansion of $\hat{\psi}_m$ has incorporated an adjustment for the unknown nuisance parameter $\lambda$. Explicit forms for $b_{ij}$, with the expansion given to one higher order, are provided in Ferguson (1991). It is also verified there that it is not possible to conclude that the asymptotic variance of $\hat{\psi}_m$ is smaller than the asymptotic variance of $\hat{\psi}$ in general.

The adjustments for the score function and the maximum likelihood estimate are determined, to first order, by the information quantity $I_{\psi\psi\psi}$; if this term is zero then $l_m$ and $l_p$ will give essentially the same results, to this order. In Cox and Reid (1991b), it is shown that $I_{\psi\psi\psi}$ determines the first order adjustment to $l_p$ in general, and that this quantity is zero if $\psi$ is a component of the expectation parameter in an exponential family, but not generally otherwise. It is also conjectured there that $l_m$ will be a substantial improvement over $l_p$ in Neyman-Scott problems, if and only if
$I_{\psi\lambda\lambda}$ is not zero. Unfortunately, the type of asymptotic expansion given above does not apply for the dimension of $\lambda$ going to $\infty$, so that we cannot conclude, for example, that $\hat{\psi}_m$ will be consistent for $\psi$ as the dimension of $\lambda$ goes to $\infty$, only that for fixed dimension it will be more nearly unbiased, in the sense described above.

In Cox and Reid (1987), stochastic expansions for $w_m(\psi) = 2l_m(\hat{\psi}_m) - l_m(\psi)$ and $w_p(\psi) = 2\{l_p(\hat{\psi}) - l_p(\psi)\}$ are given. For this summary statistic, the dominant asymptotic term is a mixture of bias and variance terms, so that the optimality described there is rather weak, in that the bias terms are reduced, but no general conclusions can be reached about the variance terms. Bartlett corrections for $w_m$ are considered in Mukerjee and Chandra (1991).

It seems likely that $r_m = \pm \sqrt{w_m}$ can be used in a Lugannani and Rice type formula, as given in (3), but experience with this is at present somewhat limited. This approach is proposed in Fraser, Reid and Wong (1991) for inference about a canonical parameter in an exponential family, and the numerical results are promising. A stochastic expansion of $r_m$ may provide some insight into this.

Two major drawbacks remain to using $l_m$, namely the lack of invariance (to $O(n^{-1})$) with respect to reparametrizations of the nuisance parameter, and the difficulty in finding an orthogonal expression of the nuisance parameter. In Cox and Reid (1989), attention is given to attempting to find the 'best' version of the orthogonal nuisance parameter $\lambda$, given that some version can be obtained by solving the orthogonality equations. The approach taken there is based on the fact that $\lambda_\psi - \hat{\lambda}$ is $O_p(n^{-1/2})$ if $\lambda$ is not orthogonal to $\psi$, but is $O_p(n^{-1})$ under orthogonality. The $O_p(n^{-1})$ term is

$$\frac{1}{n}\{\delta_{\psi} Z_{\lambda\lambda} - \frac{1}{2} \delta_{\psi} \tilde{I}_{\psi\lambda\lambda}\} + O_p(n^{-3/2}).$$

While this cannot be made zero in general, by one-to-one transformation of $\lambda$, the non-stochastic term can be made constant in $\psi$ by the choice

$$\lambda^* = \int \lambda \frac{I_{\lambda\lambda}(\psi, \chi)}{I_{\psi\lambda}(\psi, \chi)} d\chi$$

which leads to the choice $\log\{\sigma(1 - \psi^2)^{1/2}\}$ used in Cruddas, Reid and Cox (1989).

In a full exponential family, if the parameter of interest is a component of the canonical parameter, then $l_m$ is in fact invariant to one-to-one reparametrizations of the nuisance parameter $\lambda$. As shown in Barndorff-Nielsen (1983), in this case $\lambda_{\psi} = \hat{\lambda}$, so the correction term $a(\psi)$ in $l_{BN}$ is zero, and $l_m = l_{BN}$.

**Example 2.** (cont.) Since $\psi$ is a component of the canonical parameter $l_m$ and $l_{BN}$ are the same. The parameters $\lambda_i$ are components of the canonical parameter as well; the parameters that are orthogonal to $\psi$ are $p_{1i} + p_{2i}$. The modified profile likelihood written entirely in terms of the canonical parameters is $l_p(\psi) + \frac{1}{2} \log |\lambda\lambda|$, where the change of sign from (4) comes from the transformation rule for information, and the definition of the mean parameter. As in Barndorff-Nielsen (1983), we have

$$l_m(\psi) = l_{BN}(\psi) = \frac{c - b}{2} \psi - \frac{3}{2}(b + c) \log(2 + e^{\psi/2} + e^{-\psi/2}),$$
where \( b \) and \( c \) are the number of pairs \((y_{i1}, y_{i2})\) of the form \((1,0)\) and \((0,1)\) respectively, and we have assumed that there are no pairs of the form \((0,0)\) and \((1,1)\). This can be contrasted with the exact conditional likelihood for \( \psi \), obtained by conditioning on \( y_{i1} + y_{i2} \):

\[
l_c(\psi) = c\psi - (b + c) \log(1 + e^{\psi})
\]

(The conditional likelihood does not use any information from the concordant pairs \((0,0)\) and \((1,1)\), so for comparison of \( l_m \) with \( l_c \) the assumption of no concordant pairs is appropriate.) As shown in McCullagh and Tibshirani (1990), the estimating equations obtained from \( l_p \), \( l_m = l_{BN} \), and \( l_c \) are

\[
U_p(\psi) = (c - b)/2 - (1/8)(b + c)\psi + O(\psi^3)
\]
\[
U_m(\psi) = (c - b)/2 - (3/16)(b + c)\psi + O(\psi^3)
\]
\[
U_c(\psi) = (c - b)/2 - (1/4)(b + c)\psi + O(\psi^3),
\]

so that the estimate \( \hat{\psi}_m \) is more nearly consistent, as \( n \to \infty \) than is \( \hat{\psi} \). Some numerical calculations are given in Barndorff-Nielsen (1983). (The equations above correct some typographical errors in McCullagh and Tibshirani (1990).)

Example 3 (cont.) Although in this example the parameter of interest is not a component of the canonical parameter in the exponential family, it is still the case that \( \hat{\lambda}_{\psi} = \hat{\lambda}_{i} = y_{i1}/y_{i2} \), so that \( l_m = l_{BN} \) and \( l_m \) is invariant to reparametrizations of the nuisance parameter. As shown in Cox and Reid (1991b), \( \hat{\psi}_m \Rightarrow (\pi/3)\psi \), so although it is not consistent, the bias is much less than for \( \hat{\psi} \).

In Cox and Reid (1991a), approximate linearization of the orthogonality equations is considered, for cases in which the equations cannot be solved exactly. If the likelihood is presented in a non-orthogonal parameterization, say \( l(\psi, \phi) \), then \( l_m \) is given by

\[
l_m(\psi) = l(\psi, \hat{\phi}) - \frac{1}{2} \log |J_{\psi\phi}(\psi, \hat{\phi})| - \log |\frac{\partial \phi}{\partial \lambda}|
\]

using the rule for transformation of information. The third term can be approximated, near \( \hat{\psi} \), by a term of the form \((\psi - \hat{\psi})m(\psi, \hat{\phi})\); the explicit form of \( m \) is given in Cox and Reid (1991a). It is shown there that this linear approximation is sufficient to correct the bias of the score function; i.e. denoting by \( \tilde{l}_m(\psi) \) the function

\[
\tilde{l}_m(\psi) = \frac{1}{2} \log |J_{\psi\phi}| - (\psi - \hat{\psi})m(\psi, \hat{\phi})
\]

we have that \( E\tilde{l}_m''(\psi) = O(n^{-1}) \).

Direct ways of avoiding the orthogonality equations are discussed in Barndorff-Nielsen (1983, 1986) and Fraser and Reid (1989). As discussed in Section 2, this involves constructing a genuine likelihood that depends on \( \psi \) only, i.e. finding a marginal or conditional density for a statistic that is free of the nuisance parameters, at least approximately. While this is perhaps a preferable approach, it is rather difficult to implement in practice. A quite practical approach, involving resampling techniques,
to estimating adjustments to profile likelihood is given in McCullagh and Tibshirani (1990).

8. REFERENCES
