INFORMATION BIAS AND ADJUSTED PROFILE LIKELIHOOD

BY

THOMAS J. DICICCIO, MICHAEL A. MARTIN
STEVEN E. STERN and G. ALASTAIR YOUNG

TECHNICAL REPORT NO. 429
JUNE 1993

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT DMS92-04864

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Information Bias and Adjusted Profile Likelihood

By

Thomas J. DiCiccio, Michael A. Martin
Steven E. Stern and G. Alastair Young

Technical Report No. 429
June 1993

Prepared Under the Auspices

Of

National Science Foundation Grant DMS92-04864

Department of Statistics
Stanford University
Stanford, California
Information Bias and Adjusted Profile Likelihood
by Thomas J. DiCiccio, Michael A. Martin, Steven E. Stern
Department of Statistics, Stanford University, Sequoia Hall, Stanford, CA 94305, USA
and G. Alastair Young
Statistical Laboratory, University of Cambridge, Cambridge CB2 1SB, United Kingdom

SUMMARY

The bias and information bias of the profile score statistic are both typically of order $O(1)$. Several additive adjustments to the profile score statistic that reduce its bias to order $O(n^{-1})$ have been proposed in the literature. However, these adjustments do not generally reduce the order of the information bias. In this paper, an analytical formula is obtained for the information bias of a general bias-adjusted profile score statistic. This formula is used to construct further adjustments to the profile score statistic that reduce the information bias to $O(n^{-1})$. Several examples are presented to illustrate use of the formula for information bias.

Keywords: ADDITIVE ADJUSTMENT; ASYMPTOTIC EXPANSION; BIAS REDUCTION; EXPONENTIAL FAMILY; LOCATION-SCALE FAMILY; MULTIPLICATIVE ADJUSTMENT; NUISANCE PARAMETER; ORTHOGONAL PARAMETER; PARAMETERIZATION INVARIANCE; PROFILE SCORE STATISTIC
1. INTRODUCTION

Consider an observed random variable \( Y = (Y_1, \ldots, Y_n) \) having probability distribution that depends on an unknown parameter \( \theta = (\theta^1, \ldots, \theta^{p+q}) \), and let \( L(\theta) \) denote the log likelihood function for \( \theta \) based on \( Y \). Suppose \( \theta \) is partitioned in the form \( \theta = (\psi, \phi) \), where \( \psi \) is the \( p \)-dimensional parameter of interest and \( \phi \) is a nuisance parameter. Let \( \hat{\theta} = (\hat{\psi}, \hat{\phi}) \) be the overall maximum likelihood estimator of \( \theta \), and let \( \tilde{\theta}(\psi) = \{ \psi, \tilde{\phi}(\psi) \} \) be the constrained maximum likelihood estimator of \( \theta \) for a given \( \psi \). Inference for \( \psi \) is often based on the profile log likelihood function \( M(\psi) = L(\tilde{\theta}(\psi)) \). For example, \( M(\psi) \) is maximized at \( \hat{\psi} \), and the profile likelihood ratio statistic \( W(\psi) = 2 \{ M(\hat{\psi}) - M(\psi) \} \) is typically distributed as \( \chi^2_p \) to error of order \( O(n^{-1}) \).

Since the construction of \( M(\psi) \) involves estimation of the nuisance parameters, the profile log likelihood function does not behave exactly like a true log likelihood function. In particular, the profile score statistic \( \partial M(\psi)/\partial \psi \) has bias

\[
E_{\theta} \left\{ \frac{\partial M(\psi)}{\partial \psi} \right\}
\]

and information bias

\[
E_{\theta} \left\{ \frac{\partial M(\psi)}{\partial \psi^T} \frac{\partial M(\psi)}{\partial \psi} \right\} + E_{\theta} \left[ \frac{\partial}{\partial \psi^T} \left\{ \frac{\partial M(\psi)}{\partial \psi} \right\} \right]
\]

that typically do not vanish but are of order \( O(1) \).

Several authors, including Bartlett (1955), Barndorff-Nielsen and Cox (1979), Cox and Reid (1987, 1993), McCullagh and Tibshirani (1990), Barndorff-Nielsen and Chamberlin (1992), DiCiccio and Stern (1993) and Barndorff-Nielsen (1994) have proposed additive adjustments to the profile score statistic that reduce its bias to order \( O(n^{-1}) \). The bias-reducing properties of these adjustments are discussed further by Liang (1987), Levin and Kong (1990), Ferguson, Reid and Cox (1991) and Cox and Reid (1992).

Many of the proposed adjustments pertain to the case where \( \psi \) is scalar, and the adjustments are usually obtained by replacing \( M(\psi) \) with an objective function of the
form

\[ \overline{M}(\psi) = M(\psi) + B(\psi), \]

where \( B(\psi) \) is a suitably smooth function having derivatives of order \( O_p(1) \). Then, the adjustment to the profile score statistic is \( \partial B(\psi)/\partial \psi \). In this case, the estimate \( \hat{\psi} \) obtained by maximizing the adjusted profile log likelihood function \( \overline{M}(\psi) \) satisfies \( \hat{\psi} = \hat{\psi} + O_p(n^{-1}) \), and the \( \chi^2_p \) approximation to the distribution of the adjusted profile likelihood ratio statistic \( \overline{W}(\psi) = 2\{\overline{M}(\psi) - \overline{M}(\psi)\} \) has error of order \( O(n^{-1}) \). Under mild assumptions on the adjustment function \( B(\psi) \), DiCiccio and Stern (1992) showed that \( \overline{W}(\psi) \) is Bartlett correctable, and they derived a formula for the Bartlett adjustment factor.

McCullagh and Tibshirani (1990) have discussed the use of simulation to construct additive and scaling adjustments for the profile score statistic that are designed to reduce the bias and information bias, respectively. They gave an analytical expression for an additive adjustment that reduces the bias to order \( O(n^{-1}) \), and in the context of exponential families, they also provided an analytical expression for a scaling adjustment.

The present paper concerns general models. An analytical expression is given for the information bias that results when a general bias-reducing additive adjustment is applied to the profile score statistic. Although the bias is reduced to order \( O(n^{-1}) \) by additive adjustment, the information bias typically remains of order \( O(1) \). However, the analytical expression for information bias allows construction of a further adjustment that reduces the information bias to order \( O(n^{-1}) \). In addition, this expression allows investigation of the effect that previously proposed bias-reducing additive adjustments have on information bias. The formula for information bias can be used to select among orthogonal parameterizations for use with the Cox and Reid (1987) adjustment.

The analytical expression for information bias is derived in Section 2, and it is expressed in terms of expectations of the derivatives of the log likelihood function. Further adjustments to the profile score statistic that reduce information bias are also developed in Section 2. In Section 3, additive adjustments proposed by Cox and Reid (1987), Barndorff-
Nielsen (1994) and DiCiccio and Stern (1993) are examined with regard to information bias, and in each case, the relevant adjustment to reduce information bias is considered. Specific examples that facilitate comparison of these various adjustments are given in Section 4. In particular, it is shown that the bias-reducing adjustments simultaneously reduce information bias to $O(n^{-1})$ for some cases of practical interest.

2. CALCULATION OF INFORMATION BIAS

Some necessary notation is summarized at the outset. In the calculations that follow, use is made of the standard conventions for denoting arrays and sums. For applying these conventions, it is to be understood that the indices $a, b, c, \ldots$ range over $1, \ldots, p$, that the indices $i, j, k, \ldots$ range over $p + 1, \ldots, p + q$, and that the indices $r, s, t, \ldots$ range over $1, \ldots, p + q$. Differentiation is denoted by subscripts, so that $L_r(\theta) = \partial L(\theta)/\partial \theta^r$, $L_{rs}(\theta) = \partial^2 L(\theta)/\partial \theta^r \partial \theta^s$, $M_a(\psi) = \partial M(\psi)/\partial \psi^a$, $M_{ab} = \partial^2 M(\psi)/\partial \psi^a \partial \psi^b$, etc. Let $\lambda_{rs} = E_\theta \{ L_{rs}(\theta) \}$, $\lambda_{rst} = E_\theta \{ L_{rst}(\theta) \}$, $\lambda_{rs/t} = \partial \lambda_{rs}/\partial \theta^t$, etc. These quantities are assumed to be of order $O(n)$. Further, the zero-mean variables $l_r = L_r(\theta)$, $l_{rs} = L_{rs}(\theta) - \lambda_{rs}$, etc. are assumed to be of order $O_p(n^{1/2})$. These assumptions are typically satisfied in practice. Let $(\lambda^{rs})$ be the $(p + q) \times (p + q)$ matrix inverse of $(\lambda_{rs})$, and let $(\sigma_{ab})$ be the matrix inverse of $(\lambda^{ab})$, the upper left-hand $p \times p$ submatrix of $(\lambda^{rs})$. Set $\nu^{rs} = \lambda^{rs} - \lambda^{ra} \lambda^{ab} \sigma_{ab}$. The entries of the matrix $(\nu^{rs})$ are all 0, except for the lower right-hand $q \times q$ submatrix $(\nu^{ij})$, which is the matrix inverse of $(\lambda_{ij})$.

The bias of the profile score statistic is usually taken into account by making an additive adjustment $B_a(\psi)$ to $M_a(\psi)$, yielding

$$\bar{M}_a(\psi) = M_a(\psi) + B_a(\psi) \quad (a = 1, \ldots, p),$$

where the quantities $B_a(\psi) \ (a = 1, \ldots, p)$ are of order $O_p(1)$. Define the expectations $\beta_a = E_\theta \{ B_a(\psi) \}$, $\beta_{ab} = E_\theta \{ B_{ab}(\psi) \}$, etc. and put $b_a = B_a(\psi) - \beta_a$, $b_{ab} = B_{ab}(\psi) - \beta_{ab}$, etc. It is assumed that the means $\beta_a, \beta_{ab}, \ldots$ are of order $O(1)$ and that the joint cumulants
\( nb_a, nb_{ab}, l_r, l_{rs}, \ldots \) are of order \( O(n) \); specifically, \( b_a, b_{ab}, \ldots \) are assumed to be of order \( O_p(n^{-1/2}) \). These assumptions are typically satisfied in practice. In the case \( p > 1 \), the derivatives \( \psi_{a/b}(\psi) \) and \( \psi_{b/a}(\psi) \) do not necessarily coincide for \( a \neq b \), and indeed, the existence of a function \( \psi(\psi) = \partial \psi/\partial \psi^a \) \( (a = 1, \ldots, p) \) is not guaranteed. Consequently, for \( a \neq b \), \( \beta_{ab} \) and \( \beta_{ba} \) differ from \( \beta_{ba} \) and \( \beta_{ba} \) in general.

McCullagh and Tibshirani (1990) showed that

\[
E_\theta\{M_a(\psi)\} = -\rho_a(\theta) + O(n^{-1}),
\]

where

\[
\rho_a = \sigma_{ab} \lambda^{r_{ab}} \lambda^{s_{st}} \left\{ \lambda_{rs/t} - \frac{1}{2} \lambda_{rst} \right\};
\]

see also Bartlett (1955). Thus, if the adjustment terms in (4) are chosen so that

\[
\beta_a = \rho_a + O(n^{-1}) \quad (a = 1, \ldots, p),
\]

then \( E_\theta\{\psi_a(\psi)\} \) is of order \( O(n^{-1}) \). McCullagh and Tibshirani (1990) considered the case \( \psi_a(\psi) = \lambda_a(\bar{\theta}(\psi)) \); other possibilities are discussed in Section 3.

Now suppose that the adjustment terms \( \psi_a(\psi) \) \( (a = 1, \ldots, p) \) satisfy (5), and consider another adjusted profile score statistic given by

\[
\bar{M}_a(\psi) = \{\delta_a^b + C_a^b(\psi)\} \bar{M}_b(\psi) \quad (a = 1, \ldots, p),
\]

where \( \delta_a^b \) is Kronecker’s delta, \( C_a^b(\psi) = \zeta_a^b(\theta) + O_p(n^{-3/2}) \) and \( \zeta_a^b(\theta) \) is a nonrandom quantity of order \( O(n^{-1}) \). The quantity \( \bar{M}_a(\psi) \) is a scaled version of the bias-adjusted profile score statistic \( \bar{M}_a(\psi) \). Clearly the bias \( E_\theta\{\psi_a(\psi)\} \) is of order \( O(n^{-1}) \). The objective is to determine \( \zeta_a^b(\theta) \) so that the information bias

\[
\bar{\Delta}_{ab} = E_\theta\{\psi_a(\psi)\} \bar{M}_b(\psi) + E_\theta\{\psi_{a/b}(\psi)\} \quad (a, b = 1, \ldots, p)
\]

is also of order \( O(n^{-1}) \). It is assumed that \( C_{a/c}(\psi) = \zeta_{a/c}(\theta) + O_p(n^{-3/2}) \), where \( \zeta_{ac} \) is a nonrandom quantity of order \( O(n^{-1}) \).
For determining \( \zeta_a^b(\theta) \), it is useful to express \( \Delta_{ab} \) in terms of

\[
\Delta_{ab} = E_{\theta} \{ M_a(\psi)M_b(\psi) \} + E_{\theta} \{ M_{ab}(\psi) \}.
\]

By definition (6),

\[
\overline{M}_a^\dagger(\psi) = M_a(\psi) + B_a(\psi) + \zeta_a^b M_b(\psi) + O_p(n^{-1}),
\]

\[
\overline{M}_{a/b}^\dagger(\psi) = M_{ab}(\psi) + B_{a/b}(\psi) + \zeta_a^c M_c(\psi) + \zeta_a^c M_{cb}(\psi) + O_p(n^{-1}).
\]

Since \( M_a(\psi) = \sigma_{ab} \lambda^{br} l_r + O_p(1) \), it follows from these expansions that

\[
E_{\theta} \{ \overline{M}_a^\dagger(\psi) \overline{M}_b^\dagger(\psi) \} = E_{\theta} \{ M_a(\psi)M_b(\psi) \} + E_{\theta} \{ M_a(\psi)B_b(\psi) \} + E_{\theta} \{ B_a(\psi)M_b(\psi) \}
+ E_{\theta} \{ B_a(\psi)B_b(\psi) \} + E_{\theta} \{ M_a(\psi)\zeta_a^c M_c(\psi) \} + E_{\theta} \{ \zeta_a^c M_c(\psi)M_b(\psi) \} + O(n^{-1}),
\]

\[
E_{\theta} \{ \overline{M}_{a/b}^\dagger(\psi) \} = E_{\theta} \{ M_{ab}(\psi) \} + E_{\theta} \{ B_{a/b}(\psi) \} + E_{\theta} \{ \zeta_a^c M_{cb}(\psi) \} + O(n^{-1}). \tag{7}
\]

Standard calculations yield

\[
E_{\theta} \{ M_a(\psi)B_b(\psi) \} = -\rho_a \rho_b + \sigma_{ac} \rho_{b/r} \lambda^{cr} - \beta_{ba} + O(n^{-1}), \quad E_{\theta} \{ B_a(\psi)B_b(\psi) \} = \rho_a \rho_b + O(n^{-1}),
\]

\[
E_{\theta} \{ M_a(\psi)\zeta_a^c M_c(\psi) \} = -\sigma_{ac} \zeta_a^c + O(n^{-1}), \quad E_{\theta} \{ \zeta_a^c M_{cb}(\psi) \} = \zeta_a^c \sigma_{cb} + O(n^{-1}).
\]

Therefore,

\[
\overline{\Delta}_{ab} = \Delta_{ab} - \rho_a \rho_b + \rho_{a/r} \sigma_{bc} \lambda^{rc} + \sigma_{ac} \rho_{b/r} \lambda^{cr} - \beta_{ba} - \sigma_{ac} \zeta_a^c + O(n^{-1}). \tag{8}
\]

It follows from (8) that the information bias \( \overline{\Delta}_{ab} \ (a, b = 1, \ldots, p) \) is reduced to order \( O(n^{-1}) \) when

\[
\zeta_a^b = \{ \Delta_{ac} - \rho_a \rho_c + \rho_{a/r} \sigma_{cd} \lambda^{rd} + \sigma_{ad} \rho_{c/r} \lambda^{dr} - \beta_{ac} \} \lambda^{bc} + O(n^{-2})
= \Delta_{ac} \lambda^{bc} - \rho_a \rho_c \lambda^{bc} + \rho_{a/r} \lambda^{br} + \sigma_{ad} \rho_{c/r} \lambda^{bc} \lambda^{dr} - \beta_{ac} \lambda^{bc} + O(n^{-2}). \tag{9}
\]

There are several natural choices for \( C_a^b(\psi) \), notably \( C_a^b(\psi) = \zeta_a^b \{ \tilde{\theta}(\psi) \} \) and \( C_a^b(\psi) = \zeta_a^b(\tilde{\theta}) \). Formulae for \( \Delta_{ab} \) and \( \rho_{a/r} \) that facilitate evaluation of (9) are given in the Appendix.
In the case of the adjustment \( B_a(\psi) = \rho_a \{ \tilde{\theta}(\psi) \} \), it can be shown that \( \beta_{ab} = \rho_{a/r} \sigma_{bc} \lambda^{rc} + O(n^{-1}) \), and thus (9) reduces to
\[
\zeta^b_a = \left\{ \Delta_{ac} - \rho_a \rho_c + \sigma_{ad} \rho_{c/r} \lambda^{dr} \right\} \lambda^{bc} + O(n^{-2}). \tag{10}
\]
By using the formulae in the Appendix, it follows from (10) that
\[
\zeta^b_a = \sigma_{ac} \lambda^{br} \lambda^{cs} \nu^{r} \nu^{w} \left\{ \frac{1}{2} \lambda_{rst} \lambda_{uvw} - \lambda_{rst} \lambda_{uv/w} - \frac{1}{2} \lambda_{st/r} \lambda_{uvw} + \lambda_{st/r} \lambda_{uv/w} \right. \\
+ \frac{1}{2} \lambda_{rtu} \lambda_{svw} - \lambda_{rtu} \lambda_{sv/w} - \lambda_{rt/u} \lambda_{svw} - \frac{1}{2} \lambda_{tu/r} \lambda_{svw} + \lambda_{tu/r} \lambda_{sv/w} + \lambda_{ru/t} \lambda_{sv/w} \right. \\
- \sigma_{ac} \lambda^{br} \lambda^{cs} \nu^{r} \left\{ \frac{1}{2} \lambda_{rst/u} - \lambda_{rst/u} - \frac{1}{2} \lambda_{stu/r} + \lambda_{stu/r} \right\} + O(n^{-2}). \tag{11}
\]
If the additive adjustment \( B_a(\psi) \) is invariant under reparameterizations of the form \( (\psi, \phi) \to \{ \psi, \eta(\psi, \phi) \} \), then \( \beta_{ab} \) and the approximation on the right-hand side of (9) are also parameterization invariant. In particular, \( B_a(\psi) = \rho_a \{ \tilde{\theta}(\psi) \} \) is invariant under such reparameterizations, and thus the approximation on the right-hand side of (10) is invariant.

If \( \psi \) is a scalar parameter of interest, then an adjusted profile log likelihood function can be defined by
\[
\int_{\psi} \left\{ 1 + C_1(\xi) \right\} M_1(\xi) d\xi
\]
or
\[
M(\psi) + \int_{\psi} \left\{ B_1(\xi) + C_1(\xi) M_1(\xi) \right\} d\xi.
\]
The score statistics derived from these adjusted profile log likelihoods have both bias and information bias of order \( O(n^{-1}) \). Unfortunately, for vector parameters of interest, this integration approach is not generally possible because \( M^\dagger_{a/b}(\psi) \) and \( M^\dagger_{b/a}(\psi) \) are not necessarily the same to order \( O_p(n^{-1}) \) for \( a \neq b \). However, the construction of \( \zeta^b_a \) does ensure symmetry in expectation to order \( O(n^{-1}) \); that is, \( E_{\theta} \{ M^\dagger_{a/b}(\psi) \} = E_{\theta} \{ M^\dagger_{b/a}(\psi) \} + O(n^{-1}) \).

The information bias of \( M_a(\psi) \), where the additive adjustment satisfies (5), is found from (8) to be
\[
\Delta_{ab} - \rho_a \rho_b + \rho_{a/r} \sigma_{bc} \lambda^{rc} + \sigma_{ac} \rho_{b/r} \lambda^{cr} - \beta_{ba} + O(n^{-1}) \quad (a, b = 1, \ldots, p). \tag{12}
\]
Formula (12) is derived from (8) by setting $\zeta^b_a = 0$. This formula permits calculation of the information bias for various adjustments proposed in the literature.

3. INFORMATION BIAS FOR SPECIFIC ADJUSTMENTS

In this section, the information bias of the adjusted profile score statistic is calculated for additive adjustments proposed by Cox and Reid (1987), Barndorff-Nielsen (1994) and DiCiccio and Stern (1993). In each case, the function $\zeta^b_a$ is determined for use in (6) so that information bias is reduced to order $O(n^{-1})$.

In the case where $\psi$ is scalar and orthogonal to the nuisance parameters, Cox and Reid (1987) considered adjusting the profile log likelihood function according to (3) with

$$B(\psi) = -\frac{1}{2} \log \left( \frac{\det[-L_{\phi \phi}(\tilde{\theta}(\psi))]}{\det[-L_{\phi \phi}(\tilde{\theta})]} \right), \quad (13)$$

where $L_{\phi \phi}(\theta)$ is the $q \times q$ matrix of second-order partial derivatives of $L(\theta)$ taken with respect to the nuisance parameters. The associated additive adjustment $B_1(\psi) = \partial B(\psi)/\partial \psi$ to the profile score statistic satisfies (5) and has

$$\beta_{11} = -\frac{1}{2} \lambda^{ij}_1 \lambda_{11ij} + \frac{1}{2} \lambda^{ik}_1 \lambda^j_l \{\lambda_{11i} \lambda_{jkl} + \lambda_{11j} \lambda_{1kl}\} + O(n^{-1}).$$

Therefore, the information bias given by (12) is

$$\lambda^{ij}_1 \lambda_{11ij}/j - \lambda^{ik}_1 \lambda^j_l \lambda_{11i} \lambda_{jkl} + O(n^{-1}). \quad (14)$$

Although the Cox and Reid (1987) adjustment reduces the expectation of the score statistic to order $O(n^{-1})$, it does not generally reduce the information bias to order $O(n^{-1})$. However, if $\bar{M}_1^1(\psi)$ is constructed using

$$\zeta^1_1 = \lambda^{11} \lambda^{ij}_1 \lambda_{11ij}/j - \lambda^{11} \lambda^{ik}_1 \lambda^j_l \lambda_{11i} \lambda_{jkl} + O(n^{-2}), \quad (15)$$

then both the bias and information bias are of order $O(n^{-1})$. Some cases of interest where the information bias (14) is of order $O(n^{-1})$ are studied in Section 4.
For the case where $\psi$ is scalar and not necessarily orthogonal to $\phi$, Barndorff-Nielsen (1994) considered adjusting the profile log likelihood function by adding

$$B(\psi) = \int_{\psi} h\{\tilde{\theta}(\xi)\} d\xi + \frac{1}{2} \log \left( \frac{\det[-\lambda_{\phi\phi}\{\tilde{\theta}(\psi)\}]}{\det[-\lambda_{\phi\phi}(\tilde{\theta})]} \right) - \frac{1}{2} \log \left( \frac{\det[-L_{\phi\phi}\{\tilde{\theta}(\psi)\}]}{\det[-L_{\phi\phi}(\tilde{\theta})]} \right)$$

(16)

to $M(\psi)$, where

$$h(\theta) = \sigma_{11} \lambda^{11} \nu^{st}\{\lambda_{rs/t} - \frac{1}{2} \lambda_{st/r}\},$$

$\lambda_{\phi\phi}$ is the $q \times q$ submatrix of $(\lambda_{rs})$ corresponding to the nuisance parameters, and $\sigma_{11} = 1/\lambda^{11}$ by definition. Adjustment (16) is invariant under parameter transformations that preserve $\psi$, and the associated adjustment $B_1(\psi) = \partial B(\psi)/\partial \psi$ to $M_1(\psi)$ is also parameterization invariant. It can be shown that $B_1(\psi)$ satisfies (5) and that

$$\beta_{11} = -\sigma_{11}^2 \lambda^{11} \lambda^{1s} \nu^{tu}\left\{\frac{1}{2} \lambda_{rstu} - \lambda_{rt/su}\right\} + \sigma_{11}^2 \lambda^{11} \lambda^{1u} \nu^{tv} \nu^{uw}\left\{\frac{1}{2} \lambda_{rst} \lambda_{uvw} - \frac{1}{2} \lambda_{rst} \lambda_{uw/v}\right. - \lambda_{rt/s} \lambda_{uv/w} + \frac{1}{2} \lambda_{rt/s} \lambda_{uv/w} + \frac{1}{2} \lambda_{rt/s} \lambda_{sv/u} - \lambda_{tu/r} \lambda_{sv/w}\} + O(n^{-1}).$$

(17)

The information bias given by (12) is

$$\sigma_{11}^2 \lambda^{11} \lambda^{1s} \nu^{tu}\left\{\lambda_{rst/u} - \lambda_{rt/su}\right\} - \sigma_{11}^2 \lambda^{11} \lambda^{1s} \nu^{tv} \nu^{uw}\left\{\lambda_{rst} \lambda_{uvw} - \frac{1}{2} \lambda_{rst} \lambda_{uv/w} - \frac{1}{2} \lambda_{rst} \lambda_{uw/v}\right. + \frac{1}{2} \lambda_{rt/s} \lambda_{uv/w} + \lambda_{rt/s} \lambda_{sv/w} + \lambda_{rt/s} \lambda_{sv/w} - \lambda_{rt/s} \lambda_{sv/w} - \lambda_{tu/r} \lambda_{sv/w}\} + O(n^{-1}).$$

(18)

Therefore, the appropriate form of $\zeta^1$ for constructing $\overline{M}^\dagger(\psi)$ is obtained from expression (18) on multiplying by $\lambda^{11}$, which replaces the factor $\sigma_{11}^2$ with $\sigma_{11}$ and changes the error to $O(n^{-2})$. As for the Cox and Reid (1987) adjustment, Barndorff-Nielsen's (1994) adjustment reduces the bias of the profile score statistic to order $O(n^{-1})$, but it does not generally reduce the information bias to the same order.

In cases where $Y_1, \ldots, Y_n$ are independent and $\psi$ is possibly a vector parameter of interest, DiCiccio and Stern (1993) considered the additive adjustment to the profile log likelihood function

$$B(\psi) = -\frac{1}{2} \log \left( \frac{\det[-U_{\phi\phi}\{\tilde{\theta}(\psi)\}]}{\det[-U_{\phi\phi}(\tilde{\theta})]} \right) + \frac{1}{2} \log \left( \frac{\det[-L_{\phi\phi}\{\tilde{\theta}(\psi)\}]}{\det[-L_{\phi\phi}(\tilde{\theta})]} \right),$$

(19)
where \( U_{\phi}(\theta) \) is the \( q \times q \) matrix with elements

\[
U_{ij}(\theta) = \sum_{m=1}^{n} L_i(\theta; Y_m) L_j(\theta; Y_m),
\]

\( L_i(\theta; Y_m) = \partial L(\theta; Y_m)/\partial \theta^i \), and \( L(\theta; Y_m) \) is the log likelihood function for \( \theta \) based on the single observation \( Y_m \). Adjustment (19) has the advantage of not requiring calculation of expected information. In the context of an exponential family, this adjustment function coincides with the one proposed by Barndorff-Nielsen and Cox (1979) when the components of \( \psi \) are a subset of the canonical parameters. For adjustment (19), \( B_a(\psi) \) satisfies (5), and

\[
\beta_{ab} = \sigma_{ac} \sigma_{bd} \lambda^{cr} \lambda^{ds} \nu^{tv} \nu^{uw} \left\{ \frac{1}{2} \lambda_{rst} \lambda_{uvw} - \lambda_{rst} \lambda_{uv/w} + \frac{3}{2} \lambda_{rtu} \lambda_{sv/w} - 2 \lambda_{rtu} \lambda_{sv/w} ight. \\
- 2 \lambda_{rt/u} \lambda_{sw/u} + 2 \lambda_{rt/u} \lambda_{sv/w} \right\} + \sigma_{ac} \sigma_{bd} \lambda^{cr} \lambda^{ds} \nu^{tv} \nu^{uw} \left\{ \frac{1}{2} \lambda_{rstu} + \lambda_{rst/u} + \lambda_{rt,su} \right\} + O(n^{-1}),
\]

where \( \lambda_{rs,tu} = E\{\sum L_{rs}(\theta; Y_m) L_{tu}(\theta; Y_m)\} \) and \( L_{rs}^m(\theta; Y_m) = \partial^2 L(\theta; Y_m)/\partial \theta^r \partial \theta^s \). Thus, the information bias given by (12) is

\[
- \sigma_{ac} \sigma_{bd} \lambda^{cr} \lambda^{ds} \nu^{tv} \nu^{uw} \left\{ \lambda_{rst} \lambda_{uvw} - \lambda_{rst} \lambda_{uv/w} - \lambda_{rst} \lambda_{sv/w} + 2 \lambda_{rst} \lambda_{sv/w} - \lambda_{rst} \lambda_{sw/w} \right\} \\
- \sigma_{ac} \sigma_{bd} \lambda^{cr} \lambda^{ds} \nu^{tv} \nu^{uw} \lambda_{rstu} + O(n^{-1}). \tag{20}
\]

Although adjustment (19) reduces the bias of the profile score to \( O(n^{-1}) \), it does not generally reduce the information bias to the same order. However, it follows from (9) that constructing \( M_a^1(\psi) \) with

\[
\zeta_a = -\sigma_{ac} \lambda^{br} \lambda^{cs} \nu^{tv} \nu^{uw} \left\{ \lambda_{rtu} \lambda_{uvw} - \lambda_{rtu} \lambda_{uv/w} - \lambda_{rtu} \lambda_{sv/w} + 2 \lambda_{rtu} \lambda_{sv/w} - \lambda_{rtu} \lambda_{sw/w} \right\} \\
- \sigma_{ac} \lambda^{br} \lambda^{cs} \nu^{tv} \nu^{uw} \lambda_{rt,su} + O(n^{-1}) \tag{21}
\]

reduces information bias to order \( O(n^{-1}) \).
4. EXAMPLES

Example 1. Ratio of exponential means. The purpose of this example is to demonstrate how information bias can be used to choose among orthogonal parameterizations for use with the Cox and Reid (1987) adjustment (13). Suppose that $Y_m = (Y^1_m, Y^2_m)$ ($m = 1, \ldots, n$), where $Y^1_m$ and $Y^2_m$ are independent and exponentially distributed random variables having means $\phi$ and $\psi\phi$, respectively. Cox and Reid (1987) showed that $\psi$ and $\psi^{1/2}\phi$ are orthogonal parameters. Indeed, $\psi$ and $\eta$ are orthogonal if and only if $\eta = g^{-1}(\psi^{1/2}\phi)$ for a smooth invertible function $g(\eta)$. When the parameterization $(\psi, \eta)$ is used for adjustment (13), it follows from (14) that the information bias is

$$
\frac{g^{(2)}(\eta)g(\eta) - \left\{g^{(1)}(\eta)\right\}^2}{\{2\psi g^{(1)}(\eta)\}^2} + O(n^{-1}),
$$

where $g^{(k)}(\eta) = d^k g(\eta)/d\eta^k$ ($k = 1, 2$). Clearly, the information bias of $\overline{M}_1(\psi)M_1(\psi) + B_1(\psi)$ is $O(n^{-1})$ when $\eta = k_1 \log(\psi^{1/2}\phi) + k_2$ for constants $k_1, k_2$ with $k_1 \neq 0$. Cox and Reid (1993) note that for this choice of orthogonal parameters, the adjusted log profile likelihood function $M(\psi) + B(\psi)$ coincides with the logarithm of the marginal likelihood function obtained from the $F$ distribution of the pivot $\psi \sum Y^1_m / \sum Y^2_m$.

Example 2. Exponential Family. Suppose that the log likelihood function for $\theta$ based on $Y = (Y_1, \ldots, Y_n)$ is of the form

$$
L(\theta) = g^a(\psi)T_a(Y) + \phi^i T_i(Y) + D(\theta),
$$

up to an additive constant possibly depending on $Y$. In this case, $L_{ir}(\theta)$ is nonrandom, and hence $\lambda_{ir/s} = \lambda_{irs}$. For the adjustment function to the profile score statistic $B_a(\psi) = \rho_a\{\tilde{\theta}(\psi)\}$, it follows from (11) that $\zeta^b_a$ is of order $O(n^{-2})$. Therefore, this additive adjustment $B_a(\psi)$ reduces not only the bias, but also the information bias, to order $O(n^{-1})$.

When $\psi$ is scalar, adjustment (16) simplifies to

$$
B(\psi) = \int_{\bar{\psi}}^\psi \rho_1\{\tilde{\theta}(\xi)\} d\xi,
$$

10
so that \( B_1(\psi) = \rho_1 \{ \tilde{\theta}(\psi) \} \). Thus, the preceding results show that the score statistic obtained from the adjusted profile log likelihood \( M(\psi) + B(\psi) \) has bias and information bias of order \( O(n^{-1}) \).

When \( p = 1 \), it is possible to reparameterize (22) so that the nuisance parameters are orthogonal to \( \psi \) by transforming \( \phi \) to \( \eta \), where \( \eta^i = E_{\theta} \{ T_i(Y) \} \). Since \( \lambda_{11i} = 0 \) for this orthogonal parameterization, equation (15) shows that using the adjustment to the profile score statistic obtained from (13) reduces both the bias and information bias to order \( O(n^{-1}) \). Moreover, it can be shown that \( \lambda_{11i} = 0 \) for any orthogonal parameterization, and thus the reduction of both bias and information bias achieved using (13) occurs for any orthogonal parameterization.

If the observations are independent and the log likelihood function for \( \theta \) based on the single observation \( Y_m \) is

\[
L(\theta; Y_m) = g^a(\psi) T_a^m (Y_m) + \phi^iT_i^m (Y_m) + D^m(\theta),
\]

then the adjusted profile score statistic obtained using adjustment (19) has information bias

\[-\sigma_{ac} \sigma_{bd} \lambda_c \lambda_d \nu_{tu} \sum D_{rt}^m D_{su}^m + O(n^{-1}),\]

where \( D_{rs}^m = \partial^2 D^m(\theta)/\partial\theta^r \partial\theta^s \). Thus, use of adjustment (19) does not reduce the information bias to order \( O(n^{-1}) \) in general. However, if \( D_m^1(\theta) = D_m^2(\theta) \) \( (m_1, m_2 = 1, \ldots, n) \), as is the case when the observations are identically distributed, then \( D^m_{ri}(\theta) = n^{-1} \lambda_{ri} \), and use of adjustment (19) does reduce the information bias to order \( O(n^{-1}) \).

Example 3. Multiparameter Location Family. Suppose that the probability density function of \( Y \) is of the form

\[
f_Y(y; \theta) = \prod_{m=1}^{n} f^m(y_m - \theta).
\]

Then \( \lambda_{rs}, \lambda_{rst}, \text{etc.} \) do not depend on \( \theta \), and hence, \( \lambda_{rs/t}, \lambda_{rst/u}, \text{etc.} \) all vanish. In this case, \( \rho_a \) is constant, and for the additive adjustment \( B_a(\psi) = \rho_a \), the information bias is
of order $O(1)$. Indeed, it follows from (11) that $\bar{M}^\dagger(\psi)$ constructed using

$$\zeta_b^a = \frac{1}{2} \sigma_{ac} \lambda^{br} \lambda^{cs} \nu^{tv} \nu^{uw} \{ \lambda_{rst} \lambda_{uvw} + \lambda_{rtu} \lambda_{svw} \} - \frac{1}{2} \sigma_{ac} \lambda^{br} \lambda^{cs} \nu^{tv} \lambda_{rstu} + O(n^{-2}),$$

has information bias of order $O(n^{-1})$.

When the parameter of interest is scalar, adjustment (16) simplifies to

$$B(\psi) = -\frac{1}{2} \log \left( \frac{\det[-L_{\phi\bar{\phi}}(\bar{\psi}(\phi))]}{\det[-L_{\phi\bar{\phi}}(\bar{\theta})]} \right),$$

(23)

which coincides with (13), although the parameters are not generally orthogonal. It follows from (18) that the adjusted profile score statistic obtained using (23) has both bias and information bias of order $O(n^{-1})$.

When (19) is used to adjust the profile log likelihood function, the resulting score statistic has information bias

$$-\sigma_{ac} \sigma_{bd} \lambda^{cr} \lambda^{ds} \nu^{tv} \nu^{uw} \lambda_{rtu} \lambda_{svw} - \sigma_{ac} \sigma_{bd} \lambda^{cr} \lambda^{ds} \nu^{tv} \lambda_{rst},su + O(n^{-1}).$$

In particular, use of (19) does not typically reduce the information bias to order $O(n^{-1})$. However, $\bar{M}_a^\dagger(\psi)$ constructed using

$$\zeta_b^a = -\sigma_{ac} \lambda^{br} \lambda^{cs} \nu^{tv} \nu^{uw} \lambda_{rtu} \lambda_{svw} - \sigma_{ac} \lambda^{br} \lambda^{cs} \nu^{tv} \lambda_{rst},su + O(n^{-2}),$$

has information bias of order $O(n^{-1})$.

Example 4. Location-Scale Family. Suppose that the density of $Y$ can be expressed as

$$f_Y(y; \theta) = \psi^{-n} \prod_{m=1}^n f_{m} \{(y_m - \phi)/\psi\},$$

where $\phi$ is $q$-dimensional. As for the previous example, $\lambda_{rs}$, $\lambda_{rst}$, etc. do not depend on $\phi$, so that $\lambda_{rs}/i$, $\lambda_{rst}/i$, etc. all vanish. In this case, expression (11) gives

$$\zeta_1^1 = \frac{1}{2} \sigma_{11} \lambda^{1r} \lambda^{1s} \nu^{tv} \nu^{uw} \{ \lambda_{rst} \lambda_{uvw} + \lambda_{rtu} \lambda_{svw} \} - \frac{1}{2} \lambda^{1r} \nu^{su} \nu^{tv} \{ \lambda_{rst} \lambda_{uv/1} + \lambda_{rs/1} \lambda_{tuv} \} - \frac{1}{2} \sigma_{11} \lambda^{1r} \lambda^{1s} \nu^{tv} \lambda_{rstu} + \frac{1}{2} \lambda^{1r} \nu^{st} \lambda_{rst/1} + O(n^{-2}).$$
When adjustment (16) is used, it follows from (18) that the information bias is reduced to order \(O(n^{-1})\). If the underlying density functions \(f^m\) \((m = 1, \ldots, n)\) are symmetric about the origin, then \(\psi\) and \(\phi\) are orthogonal parameters and use of adjustment (13) also reduces the information bias to order \(O(n^{-1})\).

When adjustment (19) is used, expression (20) for the information bias is

\[-\sigma_2^2 \lambda^{1r} \lambda^{1s} \nu^t v^u \lambda_{rtu} \lambda_{svw} - \sigma_1^2 \lambda^{1r} \lambda^{1s} \nu^{tu} \lambda_{rt, su} + O(n^{-1}).\]

If the underlying densities are symmetric about the origin, so that \(\psi\) and \(\phi\) are orthogonal, the above expression for information bias simplifies to

\[-\lambda^{ik} \lambda^{jl} \lambda_{1ij} \lambda_{1kl} - \lambda^{ij} \lambda_{1i,1j} + O(n^{-1}).\]

As in the case of a multiparameter location family, the information bias of an adjusted profile score statistic based on adjustment (19) is typically of order \(O(1)\).

**Example 5.** Augmented Exponential Family. Suppose that the log likelihood function for \(\theta\) based on \(Y\) is of the form

\[L(\theta) = \phi^i T_i(\psi, Y) + D(\theta) + H(\psi, Y),\]

so that for fixed \(\psi\), the distribution of \(Y\) is an exponential family with canonical parameter \(\phi\). In this case, \(L_{ij}(\theta)\) is nonrandom, and hence, \(\lambda_{ij/r} = \lambda_{ijr}\), \(\lambda_{ijr/s} = \lambda_{ijrs}\), etc. For the adjustment \(B_a(\psi) = \rho_a \{\bar{\theta}(\psi)\}\), expression (11) yields

\[\begin{align*}
\s^b_a &= -\lambda^{br} \nu^{su} \nu^{tv} \left\{ \frac{1}{2} \lambda_{ars} \lambda_{tuv} - \frac{1}{2} \lambda_{as/r} \lambda_{tuv} + \lambda_{rs/t} \lambda_{auv} - \lambda_{rs/t} \lambda_{au/v} \right\} \\
&\quad + \lambda^{br} \nu^{st} \left\{ \lambda_{ars/t} - \lambda_{as/rt} \right\} + O(n^{-2}).
\end{align*}\]

(24)

For a scalar parameter of interest, adjustment (16) simplifies to

\[B(\psi) = \int_{\psi} \rho_1 \{\bar{\theta}(\xi)\} d\xi,\]

13
as in the case of a full exponential family. When this adjustment is used, the information bias is of order $O(1)$, and the appropriate $\zeta^i_1$ for constructing $\hat{M}^i_\psi(\psi)$ is given by (24).

If $E\{\partial T_i(\psi, Y)/\partial \psi^a\} = -\partial^2 D(\psi, \phi)/\partial \psi^a \partial \phi^i$, then the parameters $\psi$ and $\phi$ are orthogonal. In particular, this situation occurs when $T_i(\psi, y) = a(\psi)T_i(y)$ and $D(\psi, \phi) = a(\psi)D(\phi)$, as happens for a generalized linear model with dispersion parameter $\psi$; see McCullagh and Nelder (1989). For inference about $\psi$, formula (11) shows that $\zeta^a_\psi$ is of order $O(n^{-2})$. When $\psi$ is a scalar, the information bias obtained using either adjustments (13) or (16) is of order $O(n^{-1})$. If the observations $Y^1, \ldots, Y^m$ are independent and the log likelihood contribution from the observation $Y_m$ is

$$L(\theta; Y_m) = a(\psi)\phi^iT_i^m(Y_m) + a(\psi)D^m(\phi) + H^m(\psi, Y_m),$$

then the information bias arising from use of adjustment (19) is

$$\lambda^{ij}\lambda_{ai,bj} - \lambda^{ik}\lambda^{jl}\lambda_{aij}\lambda_{bkl} + O(n^{-1}).$$

Thus, as in the case of a full exponential family, adjustment function (19) does not typically reduce the information bias to order $O(n^{-1})$. However, unlike the case for the full exponential family, even if the observations are identically distributed, the information bias generally remains of order $O(1)$.

APPENDIX

Standard calculations show that

$$\Delta_{ab} = -\sigma_{ac}\sigma_{bd}\lambda^{cr}\lambda^{ds}\nu^{tu}\left\{\frac{1}{2}\lambda_{rstu} - \lambda_{rst/u} - \frac{1}{2}\lambda_{stu/r} - \frac{1}{2}\lambda_{rtu/s} + \lambda_{st/ru} + \lambda_{rt/su}\right\}$$

$$+ \sigma_{ac}\sigma_{bd}\lambda^{cr}\lambda^{ds}\nu^{tu}\nu^{uw}\left\{\frac{1}{2}\lambda_{rstl}\lambda_{uw} - \lambda_{rstl}\lambda_{uvw} - \frac{1}{2}\lambda_{rstl}\lambda_{uvw} - \frac{1}{2}\lambda_{rstl}\lambda_{uvw}\right.$$

$$+ \lambda_{rstl}\lambda_{uvw} + \lambda_{st/rl}\lambda_{uvw} + \frac{1}{2}\lambda_{rstl}\lambda_{suw} - \lambda_{rstl}\lambda_{suw} - \lambda_{rstl}\lambda_{suw}$$

$$- \frac{1}{2}\lambda_{rstl}\lambda_{uv/s} - \frac{1}{2}\lambda_{rstl}\lambda_{uv/s} + \lambda_{rt/u}\lambda_{uv/s} + \lambda_{rt/u}\lambda_{uv/s} + \lambda_{tu/r}\lambda_{sv/w}$$

$$+ \frac{1}{4}\lambda_{rstl}\lambda_{suw} - \frac{1}{2}\lambda_{rstl}\lambda_{suw} - \frac{1}{2}\lambda_{rstl}\lambda_{suw} + \lambda_{rstl}\lambda_{suw} + \lambda_{rstl}\lambda_{suw}\right\} + O(n^{-1})$$

14
and
\[
\rho_{a/r} = \sigma_{ab}^2 \lambda_{st/r}^2 \lambda_{uv/w} \left\{ \frac{1}{2} \lambda_{st/r} \lambda_{uv/w} - \lambda_{st/r} \lambda_{uv/w} + \frac{1}{2} \lambda_{tu/r} \lambda_{sv/w} - \lambda_{tu/r} \lambda_{sv/w} \right\}
- \sigma_{ab}^2 \lambda_{st/r}^2 \left\{ \frac{1}{2} \lambda_{st/r} + \lambda_{st/r} \right\}.
\]

REFERENCES


Levin, B. and Kong, F. (1990) Bartlett’s bias correction to the profile score function is a saddlepoint correction. Biometrika, 77, 219-221.

