Bernoulli **6**(2), 2000, 243–284

Likelihood-based inference with singular information matrix

ANDREA ROTNITZKY[1*], DAVID R. COX[2], MATTEO BOTTAI[3] and JAMES ROBINS[1,4]

[1]*Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston MA 02115, USA.* [*]*E-mail: andrea@hsph.harvard.edu*
[2]*Nuffield College, Oxford OX1 1NF, UK*
[3]*Centro Nazionale Universitario di Calcolo Elettronico, Consiglio Nazionale delle Ricerche, Via Santa Maria 36, I-56126 Pisa, Italy*
[4]*Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston MA 02115, USA*

We consider likelihood-based asymptotic inference for a p-dimensional parameter θ of an identifiable parametric model with singular information matrix of rank $p-1$ at $\theta=\theta^*$ and likelihood differentiable up to a specific order. We derive the asymptotic distribution of the likelihood ratio test statistics for the simple null hypothesis that $\theta=\theta^*$ and of the maximum likelihood estimator (MLE) of θ when $\theta=\theta^*$. We show that there exists a reparametrization such that the MLE of the last $p-1$ components of θ converges at rate $O_p(n^{-1/2})$. For the first component θ_1 of θ the rate of convergence depends on the order s of the first non-zero partial derivative of the log-likelihood with respect to θ_1 evaluated at θ^*. When s is odd the rate of convergence of the MLE of θ_1 is $O_p(n^{-1/2s})$. When s is even, the rate of convergence of the MLE of $|\theta_1-\theta_1^*|$ is $O_p(n^{-1/2s})$ and, moreover, the asymptotic distribution of the sign of the MLE of $\theta_1-\theta_1^*$ is non-standard. When $p=1$ it is determined by the sign of the sum of the residuals from the population least-squares regression of the $(s+1)$th derivative of the individual contributions to the log-likelihood on their derivatives of order s. For $p>1$, it is determined by a linear combination of the sum of residuals of a multivariate population least-squares regression involving partial and mixed derivatives of the log-likelihood of a specific order. Thus although the MLE of $|\theta_1-\theta_1^*|$ has a uniform rate of convergence of $O_p(n^{-1/2s})$, the uniform convergence rate for the MLE of θ_1 in suitable shrinking neighbourhoods of θ_1^* is only $O_p(n^{-1/(2s+2)})$.

Keywords: constraint estimation; identifiability; likelihood ratio test; non-ignorable non-response; reparametrization; rate of convergence

1. Introduction

The asymptotic distributions of the maximum likelihood estimator (MLE) and of the likelihood ratio test statistic of a simple null hypothesis in parametric models have been extensively studied when the information matrix is non-singular. In contrast, the asymptotic properties of these statistics when the information matrix is singular have been studied only in certain specific problems, but no general theory has yet been developed.

Silvey (1959) noted that in a single-parameter identifiable model the score statistic can be

1350–7265

zero for all data configurations, but did not discuss the point further. Rothenberg (1971) proved that under local identifiability the set of parameters where the score vanishes identically has Lebesgue measure zero. Cox and Hinkley (1974, p. 303) noticed that a zero score can arise in the estimation of variance components parameters, and they showed that the asymptotic distribution of the MLE of the variance components can be found after a power reparametrization. Sargan (1983) constructed identifiable simultaneous equations models with singular information matrix and derived the asymptotic distribution of specific instrumental variables estimators of the parameters of his model. Kiefer (1982) noted that in parametric mixture models that include one homogeneous distribution, the Fisher information about the mixing parameter value corresponding to the homogeneous distribution is zero. Lee and Chesher (1986) derived the asymptotic distribution of the corrected score statistic for testing homogeneity in a mixture model with a location–scale mixing distribution. Chen (1995) derived the asymptotic distribution of the MLE of the parameter indexing a finite mixture model when the true mixing parameter value corresponds to a mixing distribution with one support point. Lee (1993) calculated the asymptotic distribution of the MLEs of the parameters in a stochastic frontier function model with a singular information matrix.

Non-singular information matrices can also be encountered in parametric models that allow for non-ignorable non-response in the sense defined by Rubin (1976), also called selection bias in the econometrics literature. Lee (1981) showed that regression models with selection bias and a probit selection model (Gronau, 1974; Heckman, 1976; Nelson, 1977) can have singular information matrix. Lee and Chesher (1986) provided tests for selection bias that are asymptotically equivalent to tests based on the MLE derived in Section 2. However, they did not derive the asymptotic distribution of the MLE of the model parameters when the information matrix is singular. In particular, they did not discuss the possibility of a bimodal likelihood function and the local distributional theory when the variance of the potentially missing data is known. Copas and Li (1997) have recently carried out a largely numerical study of the performance of likelihood-based inferences in models for selection bias.

The goal of the present paper is to provide a unified theory for deriving the asymptotic distribution of the MLE and of the likelihood ratio test statistic when the information matrix has rank one less than full and the likelihood is differentiable up to a specific order. We deal only with the case of independent and identically distributed random variables, although our results can be extended to independent and non-identically distributed random variables.

The paper is organized as follows. In Section 2 we describe as a motivating example the estimation of the mean of a normal distribution under non-ignorable non-response when a parametric non-response model is assumed. In Section 3 we give some heuristics and state our results in a one-dimensional parametric model with zero Fisher information at a parameter value. The results are illustrated in a nonlinear regression model with normal errors. In Section 4 we describe our results for the multidimensional parametric models with singular information matrix of rank one less than full. Our results are applied to derive the asymptotic properties of the estimators of the example in Section 2. In Section 5 we describe some key identities from which specific asymptotic properties of the high-order derivatives of the log-likelihood follow. In Section 6 we give some final remarks.

2. A motivating example

To motivate the general discussion in the remainder of the paper, we begin with a quite informal account of a special case. The model is a simple representation of informative non-response; for similar more complicated models, see Heckman (1976) and Diggle and Kenward (1994). The example will reveal some unusual types of behaviour that it is the object of the later part of the paper to illustrate.

Suppose that Y is normally distributed with mean β and variance σ^2. There are available for study n independent individuals but for each there is the possibility that the value of Y cannot be observed. If the probability of not being able to observe Y is assumed independent of the unobserved value the analysis proceeds with just the fully observed individuals. Suppose, however, that conditionally on $Y = y$ the probability of observing y has the form

$$P_c(y;\, \alpha_0,\, \alpha_1) = \exp\{H(\alpha_0 + \alpha_1(y-\beta)/\sigma)\},$$

where (α_0, α_1) are unknown parameters and $H(\cdot)$ is a known function assumed to have its first three derivatives at α_0 non-zero. Interest may lie in small values of α_1 and in particular in testing the null hypothesis $\alpha_1 = 0$.

We thus consider two random variables (R, Y), where R is binary, taking values 0 and 1. The value of Y is observed if and only if $R = 1$. The contribution of one individual to the log-likelihood is thus

$$-r\log\sigma - r(y-\beta)^2/(2\sigma^2) + rH\{\alpha_0 + \alpha_1(y-\beta)/\sigma\} + (1-r)\log Q_c(\alpha_0,\, \alpha_1),$$

where

$$Q_c(\alpha_0,\, \alpha_1) = \mathrm{E}\{1 - P_c(Y;\, \alpha_0,\, \alpha_1)\}$$

is the marginal probability that Y is not observed. For n individuals the log-likelihood $L_n(\beta, \sigma, \alpha_0, \alpha_1)$ is the sum of n such terms.

To study behaviour for small α_1, we expand in powers of α_1. If terms are retained only up to order α_1^2 the data enter through terms in

$$n_c,\ \sum_c (y_j - \beta),\ \sum_c (y_j - \beta)^2,$$

where n_c is the number of complete observations and the summation $\sum_c$ is over the complete observations. Inference is thus to this order based on the statistics

$$\hat{p}_c = n_c/n, \qquad \bar{y}_c = \sum_c y_j/n_c, \qquad \hat{\sigma}_c^2 = \sum_c (y_j - \bar{y}_c)^2/n_c.$$

Similarly, if the expansion is carried to the term in α_1^3 the above statistics are augmented by the standardized third cumulant of the complete data, $\hat{\eta}_{c3}$.

We now consider separately the cases where σ^2 is first known and then unknown. With σ^2 known there are three unknown parameters so that in the first place we take the expansion to the term in α_1^2. Equations locally equivalent to the maximum likelihood estimating equations are then obtained by equating $\hat{p}_c$, $\bar{y}_c$, $\hat{\sigma}_c^2$ to their respective expectations.

Now

$$P_c(\alpha_0, \alpha_1) = \exp\{H(\alpha_0)\} + O(\alpha_1^2),$$

$$\mathrm{E}_c(Y) = \beta + \alpha_1 \sigma H'(\alpha_0) + O(\alpha_1^3),$$

$$\mathrm{var}_c(Y) = \sigma^2\{1 + \alpha_1^2 H''(\alpha_0)\} + O(\alpha_1^4),$$

where $P_c(\alpha_0, \alpha_1) = 1 - Q_c(\alpha_0, \alpha_1)$, $\mathrm{E}_c(Y) = \mathrm{E}(Y|R = 1)$ and $\mathrm{var}_c(Y) = \mathrm{var}(Y|R = 1)$.

It follows that the approximate estimating equations are initially of the form

$$\hat{p}_c = \exp\{H(\tilde{\alpha}_0)\},$$

$$\bar{y}_c = \tilde{\beta} + \tilde{\alpha}_1 \sigma H'(\tilde{\alpha}_0),$$

$$\hat{\sigma}_c^2 = \sigma^2(1 + \tilde{\alpha}_1^2 H''(\tilde{\alpha}_0)).$$

These equations are easily solved for the estimates; account, however, has to be taken of the restriction that $\alpha_1^2 \geqslant 0$, leading in particular to

$$\tilde{\alpha}_1^2 = \max\{(\hat{\sigma}_c^2 - \sigma^2)/H''(\tilde{\alpha}_0), 0\}.$$

Thus if interest is focused on $|\alpha_1|$, as would be the case, for example, if the sign of any non-zero value of α_1 can be regarded as known, then sampling errors in $|\tilde{\alpha}_1|$ are $O_p(n^{-1/4})$ for values of $|\alpha_1|$ that are within $O(n^{-1/4})$ of zero because the values of $\tilde{\alpha}_1^2$ are within $O_p(n^{-1/2})$ of zero. If, however, interest is in estimating β taking into account selection bias, the sign is very important, and if it is to be estimated this must be done by going to a higher-order expansion, essentially estimating $\mathrm{sgn}(\alpha_1)$ via $\mathrm{sgn}(\hat{\eta}_{c3})$, this having a probability of error of $1/2$ for these values of $|\alpha_1|$. Thus as regards the estimation of β the possibility can arise that the magnitude of the adjustment to the sample mean is reasonably well known but the direction of the adjustment essentially unknown.

Next suppose that σ^2 is unknown, giving four unknown parameters, leading directly to the use of terms up to order $O(\alpha_1^3)$ in the expansion of the log-likelihood. By the same argument as before, we obtain estimates based on $\hat{p}_c$, $\bar{y}_c$, $\hat{\sigma}_c^2$, $\hat{\eta}_{c3}$. In particular

$$\tilde{\alpha}_1^3 = \{H'''(\tilde{\alpha}_0)\}^{-1}\hat{\eta}_{c3}.$$

It follows that the sampling fluctuations in $\tilde{\alpha}_1$ are $O_p(n^{-1/6})$.

There are thus some quite striking qualitative differences in the behaviour according to whether there are three unknown parameters or four. We shall see in Section 4 the general explanation of that difference.

To study the distributional behaviour for small but non-zero values of α_1 we introduce a sequence of non-null values tending to zero as $n \to \infty$. Such sequences are totally notional and are devices for producing approximations that will be useful in appropriate circumstances.

We shall outline a number of possibilities in which

$$\alpha_1 = an^{-b},$$

where b takes values in the range $0 < b \leqslant 1/2$. This compares with the choice $b = 1/2$ in the regular Cochran–Pitman theory of asymptotic relative efficiency. The objective of the various asymptotic considerations is to highlight the quite different forms of distributions of the MLE that can occur.

First, if $b < 1/6$ the value of the third cumulant is asymptotically large compared with its standard error and the sign of α_1 is determined with high probability. Further, in the case where σ^2 is known, the atom of probability at zero in the distribution of $\tilde{\alpha}_1^2$ is also negligible.

Next, if $b = 1/6$ then asymptotically the probability that $\tilde{\alpha}_1$ has the correct sign tends to a constant bounded away from one; that is to say, the possibility that the estimate has the wrong sign cannot be ignored. Also, in the case where σ^2 is known, the atom of probability at zero in the distribution of $\tilde{\alpha}_1^2$ is negligible.

If $1/6 < b < 1/4$ then asymptotically the estimate $\tilde{\alpha}_1$ is equally likely to have either sign, and when σ^2 is known, the atom of probability at zero in the distribution of $\tilde{\alpha}_1^2$ is negligible.

If $b = 1/4$ the conclusion about $\tilde{\alpha}_1$ remains the same, but when σ^2 is known there is an atom of probability at zero in the distribution of $\tilde{\alpha}_1^2$. Finally, if $b > 1/4$ that last probability is $1/2$.

In some contexts difficulties in the asymptotic theory of maximum likelihood estimates are wholly or partially avoided by concentrating on likelihood ratio tests and on associated confidence regions, although this is much less straightforward when there are nuisance parameters and the profile likelihood is used. This is not the case here, however. Indeed, so far as inference about β is concerned, maximization over α_1 to form a profile likelihood would conceal the problems mentioned above concerning ambiguities of sign. In this paper we shall not consider the calculation of confidence regions.

Note, finally, that quite apart from the unusual distribution theory and the slow rates of convergence there is extreme sensitivity to assumptions. Thus when σ^2 is unknown any non-normality in the observed values of y is interpreted as a consequence of selection bias rather than as non-normality of the underlying distribution of Y. One of the points of our analysis is to make very explicit how this sensitivity arises.

The special features of this problem are probably best seen from the rather direct arguments summarized above. Nevertheless, to link with the general discussion in the rest of the paper and to see the essential reason for the unusual behaviour, we examine the score vector of first partial derivatives of the log-likelihood evaluated at the null point β, α_0, $\alpha_1 = 0$. The contribution from a single observation is, setting $\beta = 0$ without loss of generality,

$$ry/\sigma^2, \qquad rH'(\alpha_0) - (1-r)\mathrm{e}^{H(\alpha_0)}H'(\alpha_0)/(1-\mathrm{e}^{H(\alpha_0)}), \qquad ryH'(\alpha_0)/\sigma.$$

The key feature is that as a vector random variable this has dimension 2, because of the proportionality of the first and third components. That is, the score vector is degenerate at this particular parameter point. Equivalently, the information matrix calculated from expected second derivatives is singular at this parameter point.

3. Inferences in one-dimensional parametric models

3.1. Introduction

The asymptotic properties of the estimators in the example of Section 2 can be derived from the general asymptotic properties of the MLE in multidimensional parametric models in which the information matrix is singular and of rank one less than full. Indeed, identifiable models with zero Fisher information also exist in one-dimensional parametric models. The asymptotic derivations in the one-parameter problem are somewhat simpler than in the multiparameter problem because in the former zero information is equivalent to the vanishing of the score statistic with probability 1, while in the latter a singular information matrix is equivalent only to the existence of a linear dependence among the scores for the different parameters. Thus, in this Section we restrict attention to the one-parameter case.

Suppose that $Y_1, \ldots, Y_n$ are n independent copies of a random variable Y with density $f(y; \theta^*)$ with respect to a carrying measure, where θ^* is an unknown scalar. In Section 4 we formally state the regularity conditions assumed on $f(y; \theta)$. These essentially consist of the usual smoothness assumptions that guarantee uniqueness and consistency of the MLE and in addition the existence in a neighbourhood of θ^* of $2s+1$ derivatives with respect to θ of $\log f(Y; \theta)$ for some positive integer s with absolute values uniformly bounded by functions of Y that have finite mean. We assume, however, that exactly the first $s-1$ derivatives of $\log f(y; \theta)$ are, with probability 1, equal to 0 at $\theta = \theta^*$. That is, letting $l^{(j)}(Y; \theta)$ denote $\partial^j \log f(Y; \theta)/\partial\theta^j$, we assume that with probability 1,

$$l^{(1)}(Y; \theta^*) = l^{(2)}(Y; \theta^*) = \ldots = l^{(s-1)}(Y; \theta^*) = 0, \tag{1}$$

and with probability greater than 0,

$$l^{(s)}(Y; \theta^*) \neq 0. \tag{2}$$

Condition (1) for some $s \geqslant 2$ is equivalent to zero Fisher information at θ^*. Throughout we use $\rightsquigarrow$ to denote convergence in distribution under $\theta = \theta^*$. In a slight abuse of notation, for any pair of random variables X and W we use the identity $X = W$ to denote that X and W are equal with probability 1 if they are defined on the same probability space, otherwise to denote that they have the same distribution. In addition, we use $I(A)$ to denote the indicator of the event A, that is $I(A) = 1$ if A occurs and $I(A) = 0$ otherwise.

3.2. Informal look at inferences under $\theta = \theta^*$

An informal examination of the log-likelihood function under (1) and (2) and the regularity conditions of Section 4.2 will help provide some insight into the asymptotic properties that are stated formally later in Section 3.4. Denote the log-likelihood function $\sum l(Y_i; \theta)$ by $L_n(\theta)$, and write $L_n^{(j)}(\theta)$ for its jth derivative. A Taylor expansion of the log-likelihood around θ^* gives

$$L_n(\theta) = L_n(\theta^*) + \sum_{j=s}^{2s+1} \frac{L_n^{(j)}(\theta^*)}{j!}(\theta - \theta^*)^j + \frac{\delta_n^{(2s+1)}}{(2s+1)!}(\theta - \theta^*)^{2s+1}, \tag{3}$$

where $\delta_n^{(2s+1)} = L_n^{(2s+1)}(\bar{\theta}) - L_n^{(2s+1)}(\theta^*)$ for some $\bar{\theta}$ satisfying $|\bar{\theta} - \theta^*| < |\theta - \theta^*|$. In Section 5 we show a set of identities for the second and higher-order derivatives of the log-likelihood when (1) and (2) hold that include the following key results:

(R1.1) For $s \leqslant j \leqslant 2s-1$, $l^{(j)}(Y, \theta^*) = f^{(j)}(Y; \theta^*)/f(Y; \theta^*)$. Thus, $l^{(j)}(Y, \theta^*)$ is a mean-zero random variable. Letting $l^{[j]}(Y; \theta^*)$ denote $l^{(j)}(Y, \theta^*)/j!$, we have, in particular,

(i) $n^{-1/2} L_n^{(s)}(\theta^*)/s! = Z_0 + o_p(1)$, where $Z_0 \sim N(0, I)$ and

$$I \equiv \mathrm{E}[\{l^{[s]}(Y; \theta^*)\}^2];$$

(ii) $n^{-1/2} L_n^{(j)}(\theta^*) = O_p(1)$, $s+1 \leqslant j \leqslant 2s-1$.

(R1.2) $l^{(2s)}(Y, \theta^*) = \dfrac{f^{(2s)}(Y; \theta^*)}{f(Y; \theta^*)} - \frac{1}{2}\binom{2s}{s}\left\{\dfrac{f^{(s)}(Y; \theta^*)}{f(Y; \theta^*)}\right\}^2$. Thus,

$$n^{-1} L_n^{(2s)}(\theta^*) = -(2s)! I/2 + o_p(1).$$

(R1.3) $l^{(2s+1)}(Y, \theta^*) = \dfrac{f^{(2s+1)}(Y; \theta^*)}{f(Y; \theta^*)} - \binom{2s+1}{s}\dfrac{f^{(s)}(Y; \theta^*)}{f(Y; \theta^*)}\dfrac{f^{(s+1)}(Y; \theta^*)}{f(Y; \theta^*)}$. Thus,

$$n^{-1} L_n^{(2s+1)}(\theta^*) = -(2s+1)! C + o_p(1), \text{ where } C = \mathrm{E}\{l^{[s]}(Y; \theta^*) l^{[s+1]}(Y; \theta^*)\}.$$

Note that (R1.1) and (R1.2) are direct generalizations of familiar results and identities involving first and second log-likelihood derivatives in the theory of regular parametric models. In fact, I^{-1} coincides with the Bhattacharyya bound of order s for unbiased estimators of $(\theta - \theta^*)^s$ evaluated at $\theta = \theta^*$ (Bhattacharyya, 1946). The Bhattacharyya bound of order u for unbiased estimators of $g(\theta)$ is defined as the variance of the least-squares projection under θ of any unbiased estimator of $g(\theta)$ on the space spanned by

$$\frac{f^{(1)}(Y; \theta)}{f(Y; \theta)}, \frac{f^{(2)}(Y; \theta)}{f(Y; \theta)}, \ldots, \frac{f^{(u)}(Y; \theta)}{f(Y; \theta)}.$$

From expansion (3) these results imply that under regularity conditions on $L_n^{(2s+1)}(\theta)$, for $n^{1/2}(\theta - \theta^*)^s$ bounded,

$$L_n(\theta) - L_n(\theta^*) = G_n(\theta) + R_n, \tag{4}$$

where

$$G_n(\theta) = Z_0 n^{1/2}(\theta - \theta^*)^s - \frac{I}{2}\{n^{1/2}(\theta - \theta^*)^s\}^2,$$

and the remainder R_n converges to 0 in probability. More specifically,

$$R_n = n^{-1/(2s)}\{n^{1/(2s)}(\theta - \theta^*)\}^{s+1} T_n, \tag{5}$$

where

$$T_n = \left\{ n^{-1/2} \frac{L_n^{(s+1)}(\theta^*)}{(s+1)!} - Cn^{1/2}(\theta - \theta^*)^s + o_p(1) \right\}.$$

The expansions above show that the shape of the likelihood near θ^* depends critically on the parity of s. When s is odd, the function $G_n(\theta)$ is unimodal and has a unique global maximum at $\theta^* + n^{-1/(2s)}(Z_0/I)^{1/s}$. When s is even, $G_n(\theta)$ is symmetric around θ^*. For a positive Z_0, $G_n(\theta)$ is bimodal, it has a local minimum at θ^* and two global maxima attained at $\theta^* - n^{-1/(2s)}\ (Z_0/I)^{1/s}$ and $\theta^* + n^{-1/(2s)}(Z_0/I)^{1/s}$. For a negative Z_0, $G_n(\theta)$ is unimodal, and its global maximum is attained at θ^*. Because with probability going to 1 as $n \to \infty$ (which we will abbreviate to as with a probability tending to 1), the log-likelihood ratio $L_n(\theta) - L_n(\theta^*)$ differs from $G_n(\theta)$ locally near θ^* by a vanishing amount, we would expect, and we will later show, that under the regularity conditions of Section 4.2 the following happens:

(R2.1) The asymptotic behaviour of the MLE and of the likelihood ratio test statistic depends on the parity of s.

(R2.2) When s is odd the maximum of $L_n(\theta)$ is attained at $\theta^* + n^{-1/(2s)}\ (Z_0/I)^{1/s} + o_p(1)$.

(R2.3) When s is even, with a probability tending to 1 the maximum of $L_n(\theta)$ is attained at θ^* whenever Z_0 is negative, and at either $\tilde{\theta}_1 = \theta^* - n^{-1/(2s)}(Z_0/I)^{1/s} + o_p(1)$ or $\tilde{\theta}_2 = \theta^* + n^{-1/(2s)}(Z_0/I)^{1/s} + o_p(1)$ when Z_0 is positive.

(R2.4) Because Z_0 is a mean-zero normal random variable, when s is even the probability that the maximum is attained at θ^* converges to $1/2$ as $n \to \infty$.

(R2.5) Because $G_n(\theta)$ is symmetric around θ^* when s is even, the determination of whether the global maximum of the likelihood is attained at $\tilde{\theta}_1$ or at $\tilde{\theta}_2$ is driven by the behaviour at these two points of the remainder R_n defined in (5).

Now, since the remainder R_n is the product of $\{n^{1/(2s)}(\theta - \theta^*)\}^{s+1}$ and T_n, and this product is positive when $n^{1/(2s)}(\theta - \theta^*)$ and T_n have the same sign and is negative otherwise, then if $\hat{\theta}$ denotes the MLE of θ^*, the sign of $\hat{\theta} - \theta^*$ has to agree asymptotically with the sign of T_n. But from the results (R1.1) and (R1.3) and from the asymptotic representation of $\tilde{\theta}_1$ and $\tilde{\theta}_2$, it follows that

$$T_n = n^{-1/2}\left\{ \sum_{i=1}^{n} l^{[s+1]}(Y_i;\, \theta^*) - CI^{-1} l^{[s]}(Y_i;\, \theta^*) \right\} + o_p(1). \tag{6}$$

Thus, up to an $o_p(1)$ term, T_n is equal to the normalized sum of residuals from the population least-squares regression of $l^{[s+1]}(Y_i;\, \theta^*)$ on $l^{[s]}(Y_i;\, \theta^*)$. Because these residuals are mean-zero random variables that are uncorrelated with the regressors $l^{[s]}(Y_i;\, \theta^*)$, we conclude that $T_n \rightsquigarrow T$, a mean-zero normal variable independent of Z_0. Thus, the sign of $\hat{\theta} - \theta^*$ will be asymptotically determined by $I(T > 0)$ which is a Bernoulli random variable

with success probability $1/2$ that is statistically independent of Z_0 and hence of the absolute value of $\hat{\theta} - \theta^*$. Finally, it follows from (R2.2)–(R2.4) that $2\{L_n(\hat{\theta}) - L_n(\theta^*)\}$ converges in law to χ_1^2, a chi-squared random variable with 1 degree of freedom, when s is odd and to a mixture of a χ_1^2 random variable and 0 with mixing probabilities equal to $1/2$ when s is even. The analysis above has assumed that when s is even, $l^{(s+1)}(Y; \theta^*)$ is neither identically equal to zero nor linearly dependent with $l^{(s)}(Y; \theta^*)$, which is the case considered in this paper. See the remark at the end of Section 3.3 for some discussion on failure of this condition.

3.3. Informal look at local inferences near $\theta = \theta^*$

The discussion above has concerned behaviour at the anomalous point $\theta = \theta^*$. Somewhat similar behaviour may be expected in the neighbourhood of θ^*, and this we now explore informally. The richness of the possibilities that follow is a warning of the care needed in applications in obtaining sensible confidence regions for θ in the neighbourhood of the anomalous point θ^*. To study the behaviour of inferences locally near θ^* we consider parameters $\theta_n = \theta^* + an^{-b}$ and $\theta'_n = \theta^* - an^{-b}$ for some fixed values a and $b > 0$. Throughout, $o_{p_n}(n^{-\alpha})$ indicates a sequence of random variables that when multiplied by n^{α} converge to 0 in probability when the data are generated under θ_n. In the Appendix we provide an outline of the derivation of the following results.

Consider first $b \geqslant 1/(2s)$. Expansion (4) is valid also when the data are generated under $\theta = \theta_n$ or under $\theta = \theta'_n$, except that when $b = 1/(2s)$, $Z_0 \sim N(a^s I, I)$. Thus under regularity conditions (A1)–(A7) and (B1)–(B3) of Section 4.2, we observe the following:

(R3.1) Conclusions (R2.1)–(R2.3) and (R2.5) of the previous discussion on inferences under $\theta = \theta^*$ remain valid when the data are generated under θ_n.

(R3.2) Conclusion (R2.4) remains valid under θ_n when $b > 1/(2s)$. When $b = 1/(2s)$ and s is even, the probability that the maximum of $L_n(\theta)$ is attained at θ^* converges under θ_n to $\Phi(-a^s\sqrt{I}) < 1/2$.

(R3.3) When $b > 1/(2s)$, $2\{L_n(\hat{\theta}) - L_n(\theta^*)\}$ converges in law to the same random variable under $\theta = \theta^*$, under $\theta = \theta_n$ and under $\theta = \theta'_n$. When $b = 1/(2s)$ and s is even, $2\{L_n(\hat{\theta}) - L_n(\theta^*)\}$ converges under either θ_n or θ'_n to a mixture of the constant 0 and a non-central chi-squared random variable with 1 degree of freedom and non-centrality parameter $a^{2s}I$ with mixing probabilities $\Phi(-a^s\sqrt{I})$ and $\Phi(a^s\sqrt{I})$, respectively. When s is odd and $b = 1/(2s)$ it converges to a non-central χ_1^2 random variable with non-centrality parameter $a^{2s}I$.

(R3.4) For $b \geqslant 1/(2s)$ and s even, $2\{L_n(\hat{\theta}) - L_n(\theta_n)\}$ converges to the same random variable when the data are generated under θ_n or under θ'_n.

(R3.5) From (R3.3) and (R3.4) we have the following implications:

(i) The likelihood ratio test $2\{L_n(\hat{\theta}) - L_n(\theta^*)\}$ of the null hypothesis $H_0 : \theta = \theta^*$ has power against the alternative $H_{1n} : \theta = \theta_n$ or $H'_{1n} : \theta = \theta'_n$ that converges to its level when $b > 1/(2s)$, which is to say that the test is asymptotically completely ineffective. Its power converges to a number strictly greater than its level but bounded away from 1 when $b = 1/(2s)$.

(ii) When $b \geqslant 1/(2s)$ and s is even, the likelihood ratio test $2\{L_n(\hat{\theta}) - L_n(\theta_n)\}$ of the null hypothesis $H_{0n}: \theta = \theta_n$ has power against the alternative hypothesis $H_{1n}: \theta = \theta'_n$ that converges to its level.

(iii) Thus, from (i) and (ii), the likelihood ratio test of H_0 is sensitive to local departures of order $O(n^{-1/(2s)})$ but not of order $O(n^{-1/2})$. However, when s is even as the sample size n converges to ∞, the data provide essentially no indication of the directionality of departures of order $O(n^{-1/(2s)})$.

Now consider $1/(2s+2) < b < 1/(2s)$. Under the regularity conditions stated in the Appendix, we have that

$$L_n(\theta^*) - L_n(\theta_n) \text{ converges in probability to } -\infty \text{ under } \theta_n. \tag{7}$$

Furthermore, when s is odd,

$$L_n(\theta'_n) - L_n(\theta_n) \text{ converges in probability to } +\infty \text{ under } \theta'_n. \tag{8}$$

However, for any s and for θ satisfying

$$\sqrt{n}\{(\theta - \theta^*)^s - (\theta_n - \theta^*)^s\} = O(1), \tag{9}$$

the log-likelihood function satisfies

$$L_n(\theta) - L_n(\theta_n) = \tilde{G}_n(\theta) + \tilde{R}_n, \tag{10}$$

where

$$\tilde{G}_n(\theta) = Z_0 n^{1/2}\{(\theta - \theta^*)^s - (\theta_n - \theta^*)^s\} - \frac{I}{2}[n^{1/2}\{(\theta - \theta^*)^s - (\theta_n - \theta^*)^s\}]^2, \tag{11}$$

and $\tilde{R}_n = n^{1/2}\{(\theta - \theta^*)^{s+1} - (\theta_n - \theta^*)^{s+1}\}\tilde{T}_n$, with $\tilde{T}_n = Z_1 - Cn^{1/2}\{(\theta - \theta^*)^s - (\theta_n - \theta^*)^s\} + o_{p_n}(1)$.

Here I and C are defined as before and (Z_0, Z_1) is a bivariate mean-zero normal random vector with $\mathrm{var}(Z_0) = I$, $\mathrm{cov}(Z_0, Z_1) = C$ and $\mathrm{var}(Z_1) = J$, where $J = \mathrm{E}[\{l^{[s+1]}(Y; \theta^*)\}^2]$. Now, when s is odd and (9) is true, $\tilde{G}_n(\theta)$ has a unique maximum attained at $\theta^* + \{(\theta_n - \theta^*)^s + n^{-1/2}Z_0/I\}^{1/s}$. When s is even, for values of θ satisfying (9) and such that $\mathrm{sgn}(\theta - \theta^*)$ is constant, $\tilde{G}_n(\theta)$ is concave. Furthermore, $\tilde{G}_n(\theta)$ is symmetric around θ^* and has two global maxima attained at $\theta^* + \{(\theta_n - \theta^*)^s + n^{-1/2}Z_0/I\}^{1/s}$ and $\theta^* - \{(\theta_n - \theta^*)^s + n^{-1/2}Z_0/I\}^{1/s}$. Also, because $n^{1/2}\{(\theta - \theta^*)^{s+1} - (\theta_n - \theta^*)^{s+1}\} = o(1)$ when (9) holds, $\tilde{R}_n = o_{p_n}(1)$. This suggests the following results.

(R4.1) From (7),

$$L_n(\hat{\theta}) - L_n(\theta^*) \text{ converges to } +\infty \text{ under } \theta_n \tag{12}$$

and therefore the asymptotic distribution of the MLE no longer has an atom of probability at θ^*.

(R4.2) When s is odd, $L_n(\theta)$ has a unique maximum at $\hat{\theta}$ satisfying

$$\begin{aligned}\hat{\theta} &= \theta^* + \{(\theta_n - \theta^*)^s + n^{-1/2}Z_0/I + o_{p_n}(n^{-1/2})\}^{1/s} \\ &= \theta_n + n^{b(s-1)-1/2}a^{s-1}Z_0/(sI) + o_{p_n}(n^{b(s-1)-1/2}).\end{aligned} \tag{13}$$

(R4.3) When s is even, the MLE $\hat{\theta}$ of θ is equal to $\tilde{\theta}_1$ or $\tilde{\theta}_2$ satisfying, for $j = 1, 2$,

$$\tilde{\theta}_j - \theta^* = (-1)^{j+1}\mathrm{sgn}(\theta_n - \theta^*)\{(\theta_n - \theta^*)^s + n^{-1/2}Z_0/I + o_{p_n}(n^{-1/2})\}^{1/s}. \tag{14}$$

(R4.4) Determining at which of $\tilde{\theta}_1$ or $\tilde{\theta}_2$ $L_n(\theta)$ attains its global maximum is driven by the behaviour of the remainder $\tilde{R}_n$ evaluated at these two points. Thus, as argued earlier for the case of inferences under $\theta = \theta^*$, the sign of the MLE has to agree asymptotically with the sign of $\tilde{T}_n$ evaluated at either point. But at either point, $\tilde{T}_n$ satisfies $\tilde{T}_n = Z_1 - CI^{-1}Z_0 + o_{p_n}(1)$. Thus, since $Z_1 - CI^{-1}Z_0$ is a mean-zero normal random variable, we conclude that the probability that the sign of $\hat{\theta} - \theta^*$ is the same as the sign of $\theta_n - \theta^*$ converges to $1/2$.

(R4.5) Evaluating $L_n(\theta)$ at $\tilde{\theta}_1$ and at $\tilde{\theta}_2$, we obtain

$$L_n(\hat{\theta}) - L_n(\theta_n) = \tfrac{1}{2}\{Z_0/\sqrt{I}\}^2 + o_{p_n}(1). \tag{15}$$

(R4.6) Because when s is even, $\tilde{G}_n(\theta'_n) = 0$, then $L_n(\theta'_n) - L_n(\theta_n) = o_{p_n}(1)$. Thus, (15) and (12) also hold when the data are generated under θ'_n.

(R4.7) From (R4.1), (R4.5), (R4.6) and equation (8) we conclude that

(i) The likelihood ratio test $2\{L_n(\hat{\theta}) - L_n(\theta^*)\}$ of the null hypothesis $H_0: \theta = \theta^*$ has power that converges to 1 for detecting the alternative hypothesis $H_{1n}: \theta = \theta_n$ or $H'_{1n}: \theta = \theta'_n$.

(ii) The likelihood ratio test $2\{L_n(\hat{\theta}) - L_n(\theta_n)\}$ of the null hypothesis $H_{0n}: \theta = \theta_n$ has power against the alternative hypothesis $H_{1n}: \theta = \theta'_n$ that converges to its level when s is even. When s is odd the power converges to 1.

(iii) From (i) and (ii) we conclude that the likelihood ratio test of H_0 has power converging to 1 for detecting local departures of order $O(n^{-b})$. Nevertheless, when s is even the directionality of the departure is left inconclusive.

Now consider $b = 1/(2s + 2)$. Under the regularity conditions stated in the Appendix, (7) and (8) remain valid. However, when (9) holds,

$$L_n(\theta) - L_n(\theta_n) = \bar{G}_n(\theta) + \bar{R}_n(\theta), \tag{16}$$

where

$$\bar{G}_n(\theta) = \{Z_0 + d_s(\theta)2a^{s+1}C\}n^{1/2}\{(\theta - \theta^*)^s - (\theta_n - \theta^*)^s\}$$

$$-\frac{I}{2}[n^{1/2}\{(\theta - \theta^*)^s - (\theta_n - \theta^*)^s\}]^2,$$

$$d_s(\theta) = \tfrac{1}{4}\{1 + (-1)^s\}|\mathrm{sgn}(\theta - \theta^*) - \mathrm{sgn}(\theta_n - \theta^*)|,$$

and

$$\bar{R}_n(\theta) = -d_s(\theta)2a^{s+1}\{Z_1 + a^{s+1}J\} + o_{p_n}(1).$$

When s is odd, $d_s(\theta) = 0$ and therefore for values of θ satisfying (9),

$$L_n(\theta) - L_n(\theta_n) = \tilde{G}_n(\theta) + o_{p_n}(1), \tag{17}$$

where $\tilde{G}_n(\theta)$ is defined in (11). This suggests (and it is shown in the Appendix under regularity conditions) that (R4.2) for the case $1/(2s+2) < b < 1/(2s)$ remains valid when $b = 1/(2s+2)$.

When s is even, equation (16) describes the shape of the log-likelihood function for values of θ satisfying (9) and the role of the constant C in determining its shape, as well as suggesting the behaviour of likelihood-based inferences. Specifically, we have the following results:

(R5.1) For values of θ satisfying (9) such that $\operatorname{sgn}(\theta - \theta^*)$ is constant, the function $\bar{G}_n(\theta)$ is concave. The two local maxima of $\bar{G}_n(\theta)$ are attained at the points

$$\theta^* + \operatorname{sgn}(\theta_n - \theta^*)\{(\theta_n - \theta^*)^s + n^{-1/2} Z_0 / I\}^{1/s}$$

and

$$\theta^* - \operatorname{sgn}(\theta_n - \theta^*)[(\theta_n - \theta^*)^s + n^{-1/2}\{Z_0 + 2a^{s+1} C\}/I]^{1/s}$$

.

(R5.2) From (R5.1) we would expect (and we show in the Appendix) that under regularity conditions the maximum of $L_n(\theta)$ is attained at one of the points $\tilde{\theta}_1$ and $\tilde{\theta}_2$ satisfying

$$\begin{aligned}\tilde{\theta}_1 &= \theta^* + \operatorname{sgn}(\theta_n - \theta^*)\{(\theta_n - \theta^*)^s + n^{-1/2} Z_0 / I + o_{p_n}(n^{-1/2})\}^{1/s} \\ &= \theta_n + n^{-1/(s+1)} a^{s-1} Z_0/(sI) + o_{p_n}(n^{-1/(s+1)}) \qquad (18)\end{aligned}$$

and

$$\begin{aligned}\tilde{\theta}_2 &= \theta^* - \operatorname{sgn}(\theta_n - \theta^*)[(\theta_n - \theta^*)^s + n^{-1/2}\{Z_0 + 2a^{s+1} C\}/I + o_{p_n}(n^{-1/2})]^{1/s} \\ &= 2\theta^* - \theta_n - n^{-1/(s+1)} a^{s-1}\{Z_0 + 2a^{s+1} C\}/(sI) + o_{p_n}(n^{-1/(s+1)}). \qquad (19)\end{aligned}$$

(R5.3) By (16), (18) and (19),

$$L_n(\tilde{\theta}_1) - L_n(\tilde{\theta}_2) = 2(Z_{1.0} + \sigma^2_{1.0}) + o_{p_n}(1), \qquad (20)$$

where

$$Z_{1.0} = a^{s+1}(Z_1 - CI^{-1} Z_0) \quad \text{and} \quad \sigma^2_{1.0} = a^{2s+2}(J - C^2 I^{-1}).$$

Now, since (7) holds when $b = 1/(2s+2)$, (R4.1) for the case $1/(2s+2) < b < 1/(2s)$ remains valid. From (20) we conclude that, asymptotically, the global maximum of $L_n(\theta)$ is attained at $\tilde{\theta}_1$ whenever $Z_{1.0} + \sigma^2_{1.0} > 0$ and at $\tilde{\theta}_2$ otherwise. Since $Z_{1.0}$ is a mean-zero normal random variable with variance equal to $\sigma^2_{1.0}$, the probability that the MLE $\hat{\theta}$ coincides with $\tilde{\theta}_1$ (and therefore that $\operatorname{sgn}(\hat{\theta} - \theta^*) = \operatorname{sgn}(\theta_n - \theta^*)$) converges to $\Phi(\sqrt{\sigma^2_{1.0}}) > 1/2$.

(R5.4) The distribution of $Z_{1.0} + \sigma^2_{1.0}$ is the same as the limit law under θ_n of

$$n^{-1/2} \sum \{l^{[s+1]}(Y_i; \theta^*) - \beta l^{[s]}(Y_i; \theta^*)\}, \qquad (21)$$

where $\beta = \operatorname{cov}\{l^{[s+1]}(Y; \theta^*),\ l^{[s]}(Y; \theta^*)\}\operatorname{var}\{l^{[s]}(Y; \theta^*)\}^{-1}$. Thus, the sign of

$(\hat{\theta} - \theta^*)$ is asymptotically determined by (21). The sign of the random variable $l^{[s+1]}(Y; \theta^*) - \beta l^{[s]}(Y; \theta^*)$ is interpreted as the effective score for estimating $\text{sgn}(\theta - \theta^*)$ after taking into account the estimation of $|\theta - \theta^*|$.

(R5.5) Suppose now that $C = 0$. Then $\bar{G}_n(\theta)$ is symmetric around θ^*. Thus, for values of $\theta^{(j)}$, $j = 1, 2$, satisfying (9) and such that $\theta^{(1)} - \theta^* = -(\theta^{(2)} - \theta^*)$ with, say, $\text{sgn}(\theta^{(1)} - \theta^*) = \text{sgn}(\theta_n - \theta^*)$, $L_n(\theta^{(1)})$ differs, up to an $o_{p_n}(1)$ term, from $L_n(\theta^{(2)})$ by the value $2a^{s+1}\{Z_1 + a^{s+1}J\}$ independent of θ. But since, by (R5.3), asymptotically $\text{sgn}(\hat{\theta} - \theta^*) = \text{sgn}(\theta_n - \theta^*)$ if and only if $2a^{s+1}\{Z_1 + a^{s+1}J\} > 0$, we conclude that for any pair of points equidistant from θ^* satisfying (9) the likelihood will tend to be greater at the point whose difference from θ^* has the same sign as the difference of the MLE from θ^*. Thus, likelihood-based confidence regions for θ when the data are generated under θ_n will, with high probability for large samples, be comprised of two disjoint intervals located at each side of θ^*, with the interval located on the same side as the MLE having the largest length.

(R5.6) If $C \neq 0$, after some algebra it can be shown that $L_n(\theta^{(1)}) - L_n(\theta^{(2)})$ is, up to an $o_{p_n}(1)$ term, equal to

$$\frac{2a^{s+1}C}{I}[a^{s+1}C - In^{1/2}\{(\theta^{(2)} - \theta^*)^s - (\hat{\theta} - \theta^*)^s\}] + 2(Z_{1.0} + \sigma^2_{1.0})$$

if $Z_{1.0} + \sigma^2_{1.0} > 0$ or equivalently if $\text{sgn}(\hat{\theta} - \theta^*) = \text{sgn}(\theta_n - \theta^*)$, and it is equal to

$$\frac{2a^{s+1}C}{I}[-a^{s+1}C - In^{1/2}\{(\theta^{(2)} - \theta^*)^s - (\hat{\theta} - \theta^*)^s\}] + 2(Z_{1.0} + \sigma^2_{1.0})$$

if $Z_{1.0} + \sigma^2_{1.0} < 0$ or equivalently if $\text{sgn}(\hat{\theta} - \theta^*) \neq \text{sgn}(\theta_n - \theta^*)$. Thus, for values of $\theta^{(j)}$, $j = 1, 2$, equidistant from θ^* and such that $|\theta^{(j)} - \theta^*|$ differs by a small amount from $|\hat{\theta} - \theta^*|$, the likelihood will be larger at the value $\theta^{(j)}$ located on the same side of θ^* as the MLE. However, for moderate and large values of $|(\theta^{(j)} - \theta^*)^s - (\hat{\theta} - \theta^*)^s|$ the sign of the difference $L_n(\theta^{(1)}) - L_n(\theta^{(2)})$ will be essentially determined by the sign of the function

$$-2a^{s+1}Cn^{1/2}\{(\theta^{(2)} - \theta^*)^s - (\hat{\theta} - \theta^*)^s\}$$

and will therefore depend on the sign of aC or equivalently of $\theta_n C$. In particular, if C has sign opposite to the sign of θ_n, so that $aC < 0$, then as the parameter points move away from θ^* the likelihood function will tend to decrease more rapidly on the side of θ^* opposite to where θ_n is located. The contrary will occur when $aC > 0$. Thus, when the data are generated under θ_n, then, with high probability for large samples, the relative length of the intervals comprising a likelihood-based confidence region for θ will depend on the sign of $\theta_n C$.

(R5.7) By (16), (18) and (19),

$$\begin{aligned} L_n(\hat{\theta}) - L_n(\theta_n) &= \bar{G}_n(\tilde{\theta}_1)I(Z_{1.0} + \sigma^2_{1.0} > 0) \\ &\quad + \{\bar{G}_n(\tilde{\theta}_2) + \bar{R}_n(\tilde{\theta}_2)\}I(Z_{1.0} + \sigma^2_{1.0} < 0) + o_{p_n}(1) \\ &= \tfrac{1}{2}(Z_0/\sqrt{I})^2 - 2(Z_{1.0} + \sigma^2_{1.0})I(Z_{1.0} + \sigma^2_{1.0} < 0) + o_{p_n}(1). \end{aligned} \tag{22}$$

Thus, the likelihood ratio test $2\{L_n(\hat{\theta}) - L_n(\theta_n)\}$ of the null hypothesis $H_{0n}: \theta = \theta_n$ converges to the sum of a χ^2_1 random variable and an independent truncated positive normal random variable.

(R5.8) Equation (22) is valid also when the data are generated under $\theta = \theta'_n$, except that $Z_0 \sim N(-2a^{s+1}C, I)$ and $Z_{1.0} \sim N(-2\sigma^2_{1.0}, \sigma^2_{1.0})$. Thus, the likelihood ratio test $2\{L_n(\hat{\theta}) - L_n(\theta_n)\}$ of the null hypothesis $H_{0n}: \theta = \theta_n$ has power against the alternative $H_{1n}: \theta = \theta'_n$ that converges to a value strictly greater than its level but bounded away from 1.

(R5.9) From (R5.7) and (R5.8), the likelihood ratio test of $H_0: \theta = \theta^*$ has power converging to 1 for detecting local departures of order $O(n^{-1/(2s+2)})$. Nevertheless, the directionality of the departure is only correctly determined with probability that converges to a number bounded away from 1.

(R5.10) From (22), the likelihood ratio test $2\{L_n(\hat{\theta}) - L_n(\theta_n)\}$ of the null hypothesis $H_{0n}: \theta = \theta_n$ converges in law under θ_n to a random variable that is stochastically larger than the χ^2_1 random variable. This implies that likelihood-based confidence regions computed using the $1-\alpha$ critical point of the χ^2_1 distribution will not have uniform asymptotic coverage equal to $1-\alpha$.

Finally consider $b < 1/(2s+2)$. Under the regularity conditions of the Appendix, (7) holds and therefore the likelihood ratio test of the hypothesis $H_0: \theta = \theta^*$ has power converging to 1 for detecting the alternative hypotheses $H_{1n}: \theta = \theta_n$ and $H'_{1n}: \theta = \theta_n$. Furthermore, (8) holds for all values of s. Thus, $2\{L_n(\hat{\theta}) - L_n(\theta_n)\}$ converges in probability to $+\infty$ under θ'_n. Thus, the likelihood ratio test of the hypothesis $H_{0n}: \theta = \theta_n$ has power converging to 1 for detecting the alternative hypothesis $H_{1n}: \theta = \theta'_n$. We conclude that departures from θ^* of order $O(n^{-b})$ are detected with a probability tending to 1 and the directionality of the departure is firmly determined.

Figure 1 illustrates the variety of likelihood functions that arise. More detailed properties are summarized in the previous results.

Remark. When $l^{(s+1)}(Y; \theta^*)$ vanishes identically but $l^{(s+3)}(Y; \theta^*)$ is not zero, the determination of the sign of $(\hat{\theta} - \theta^*)$ under θ^* is driven asymptotically by higher-order terms of the log-likelihood expansion. Specifically, it can be shown that under θ^*, $\operatorname{sgn}(\hat{\theta} - \theta^*)$ is asymptotically determined by the sign of the sum of residuals from the population least-squares regression of $l^{[s+3]}(Y_i; \theta^*)$ on $l^{[s]}(Y_i; \theta^*)$. Also, in this case, Z_1, J and C are all equal to 0. Then, from (16), $L_n(\theta'_n) - L_n(\theta_n) = o_{p_n}(1)$ when s is even and $b = 1/(2s+2)$. Thus, the likelihood ratio test of $H_{0n}: \theta = \theta_n$ has power converging to its level for detecting the alternative $H_{1n}: \theta = \theta'_n$. A similar situation occurs in the example of Section 2 if $H'''(\alpha_0) = 0$ and the fifth derivative $H^{v}(\alpha_0)$ is not equal to 0. When $l^{(s+1)}(Y_i; \theta^*)$ does not vanish identically but is equal to $Kl^{(s)}(Y_i; \theta^*)$ for some non-zero constant K, the reparametrization $\psi = \theta - \theta^* + K\{s(s+1)\}^{-1}(\theta - \theta^*)^2$ yields a model with 1st, ..., $(s-1)$th and $(s+1)$th log-likelihood derivatives that vanish identically at $\psi^* = 0$, so that the above remarks hold for the estimation of the sign of ψ. Interestingly, in this case the likelihood ratio test of $H_{0n}: \theta = \theta_n$ when s is even and $b = 1/(2s+2)$ has power for detecting the alternative $H_{1n}: \theta = \theta'_n$ converging to a number strictly greater than its level

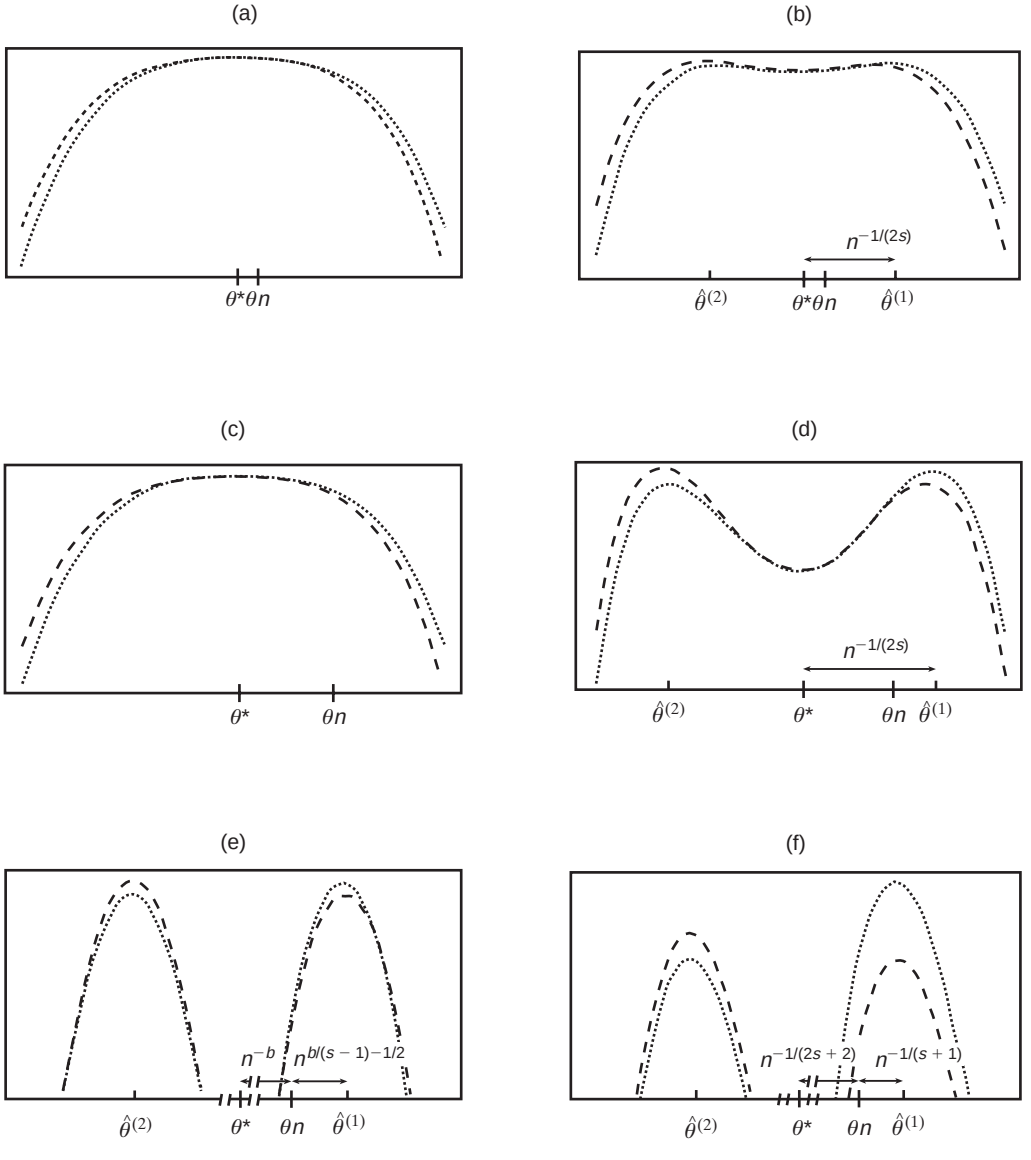

Figure 1. Schematic forms of $L_n(\theta) - L_n(\hat{\theta})$ for s even; data generated under $\theta_n = \theta^* + an^{-b}$. (a) $b > 1/(2s)$: two log-likelihoods with $\hat{\theta} = \theta^*$, asymptotic probability $1/2$. (b) $b > 1/(2s)$: two bimodal log-likelihoods with $\hat{\theta} = \hat{\theta}^{(1)}$ or $\hat{\theta} = \hat{\theta}^{(2)}$, asymptotic probability $1/2$. (c) $b = 1/(2s)$: two log-likelihoods with $\hat{\theta} = \theta^*$, asymptotic probability less than $1/2$. (d) $b = 1/(2s)$: two bimodal log-likelihoods with $\hat{\theta} = \hat{\theta}^{(1)}$ or $\hat{\theta} = \hat{\theta}^{(2)}$, asymptotic probability greater than $1/2$. (e) $1/(2s+2) < b < 1/(2s)$: dotted curve, global maximum at $\hat{\theta}^{(1)}$; dashed curve, global maximum at $\hat{\theta}^{(2)}$; asymptotic probability of both types $1/2$. (f) $b = 1/(2s+2)$: as for (e) except that the asymptotic probability of the dotted curve is greater than $1/2$.

even though the difference $L_n(\tilde{\theta}_1) - L_n(\tilde{\theta}_2) = o_{p_n}(1)$. Thus, the likelihood ratio test of the composite null hypothesis $H_0' : \theta > \theta^*$ remains asymptotically ineffective for detecting the alternative hypothesis $H_1' : \theta < \theta^*$. Further terms of the log-likelihood expansion are required if both $l^{(s+1)}(Y_i; \theta^*)$ and $l^{(s+3)}(Y_i; \theta^*)$ vanish identically. Alternatively, a second reparametrization is needed if $l^{(s+1)}(Y_i; \theta^*)$ is linearly dependent with $l^{(s)}(Y_i; \theta^*)$ and the $(s+3)$th and sth derivatives of the log-likelihood under the initial reparametrization are also linearly dependent.

3.4. Asymptotic properties under $\theta = \theta^*$

The following two theorems establish the asymptotic distribution of the MLE of θ and of the likelihood ratio test statistic when $\theta = \theta^*$ for odd and even values of s. They are special cases of the Theorems 3 and 4 stated in Section 4. We nevertheless state them here for the sake of completeness.

Theorem 1. *Under regularity conditions (A1)–(A7), (B1) and (B2) of Section 4.2, when s is odd: (a) the MLE $\hat{\theta}$ of θ exists when $\theta = \theta^*$, it is unique with a probability tending to 1 and it is a consistent estimator of θ when $\theta = \theta^*$; (b) $n^{1/(2s)}(\hat{\theta} - \theta^*) \rightsquigarrow Z^{1/s}$, where $Z \sim N(0, I^{-1})$; and (c) $2\{L_n(\hat{\theta}) - L_n(\theta^*)\} \rightsquigarrow \chi_1^2$.*

Theorem 2. *Under regularity conditions (A1)–(A7) and (B1)–(B3) of Section 4.2, when s is even: (a) the MLE $\hat{\theta}$ of θ exists when $\theta = \theta^*$, it is unique with a probability tending to 1 and it is a consistent estimator of θ when $\theta = \theta^*$; (b) $n^{1/(2s)}(\hat{\theta} - \theta^*) \rightsquigarrow (-1)^B Z^{1/s} I(Z > 0)$, where B is a Bernoulli random variable with success probability equal to $1/2$, and $Z \sim N(0, I^{-1})$ independent of B; and (c) $2\{L_n(\hat{\theta}) - L_n(\theta^*)\} \rightsquigarrow Z^{*2} I(Z^* > 0)$, where $Z^* \sim N(0, 1)$.*

Theorems 1 and 2 imply that when the Fisher information is zero at a parameter θ^* then, under suitable regularity conditions, there exists s such that $n^{1/2}(\hat{\theta} - \theta^*)^s$ is asymptotically normally distributed with variance that attains the Bhattacharyya bound of order s for unbiased estimators of $(\theta - \theta^*)^s$ evaluated at θ^*.

In the Appendix we give the proofs of Theorems 1 and 2. These proofs essentially rely on a first-order expansion of the score function in order to determine the asymptotic distribution of the roots of the score equation, and then on the expansion of the remainder R_n to determine at which of these roots the likelihood is maximized. In particular, the proofs show that the difference between the likelihood ratio test statistic and its limiting random variable is of order $O_p(n^{-1/(2s)})$ and, when s is even, this difference is positive. The difference between the respective cumulative distribution functions is of order $O(n^{-1/(2s)})$ when s is even, but it is of order $O(n^{-1/s})$ when s is odd (Cox and Reid 1987a).

Some but not all of the above results when the data are generated under $\theta = \theta^*$ can be obtained by reparametrization. If s is odd, one can use the one-to-one transformation $\lambda = \theta^* + (\theta - \theta^*)^s$ to show that the model admits a regular parametrization (Bickel *et al.* 1993, Section 2.1, Proposition 1) and hence the results of Theorem 1 follow. In fact, this

argument shows that the model is locally asymptotically normal (LAN) at $\lambda = \theta^*$ with normalizing constant of order $O(n^{1/2})$ (Ibragimov and Has'minskii 1981, Section 2.1, Theorem II.1.2; Bickel *et al.* 1993, Section 2.1. Proposition 2). In particular, the LAN property implies that the optimal rate for estimating λ under $\lambda = \theta^*$ is no better than $O_p(n^{-1/2})$, and thus the optimal rate for estimating for θ is no better than $O_p(n^{-1/(2s)})$ (Ibragimov and Has'minskii 1981, Section 2.9, Lemma 9.1). Interestingly, the second derivative of the log-likelihood in the reparametrized model does not exist at $\lambda = 0$ unless all of the derivatives of $\log f(y; \theta)$ at $\theta = \theta^*$ of order $s+1$ to $2s-1$ are 0. When s is even, the transformation $\lambda = \theta^* + (\theta - \theta^*)^s$ in effect yields inference only about $|\theta - \theta^*|$.

3.5. Example

Suppose that $Y_i = (W_i, X_i)$, $i = 1, \ldots, n$, are independent bivariate random variables. Suppose that the marginal law of X_i is known and that, conditional on X_i,

$$W_i = \exp(-\theta X_i) - \sum_{k=0}^{s-1} \frac{(-1)^k}{k!} \theta^k X_i^k + \varepsilon_i, \tag{23}$$

with $\varepsilon_i \sim N(0, \sigma^2)$. Suppose that σ^2 is known but θ is unknown. A simple calculation shows that the first $s-1$ derivatives of the log-likelihood with respect to θ evaluated at $\theta = 0$ are identically equal to zero. Furthermore, at $\theta = 0$, $l^{[s]}(Y_i; \theta^*) = (s!)^{-1}\sigma^{-2}(-1)^s \varepsilon_i X_i^s$ and $l^{[s+1]}(Y_i; \theta^*) = (s!)^{-1}\sigma^{-2}(-1)^{s+1}\varepsilon_i X_i^{s+1}$. Thus when s is odd, $n^{1/(2s)}(\hat{\theta} - 0) \rightsquigarrow Z^{1/s}$, and when s is even, $n^{1/(2s)}(\hat{\theta} - 0) \rightsquigarrow (-1)^B Z^{1/s} I(Z > 0)$, where $Z \sim N(0, I^{-1})$ with $I = \{(\sigma s!)^{-2}\ \mathrm{E}(X_i^{2s})\}^{-1}$. The random variable B is Bernoulli with success probability $1/2$ independent of Z. As noted before, the distribution of B follows because it is the limiting distribution of the sign of the sequence T_n in (6). In this example, T_n is equal to

$$\sum_i^n \varepsilon_i \left\{ \frac{(-1)^{s+1} X_i^{s+1}}{(s+1)!} - \gamma \frac{(-1)^s X_i^s}{s!} \right\}, \tag{24}$$

where $\gamma = -(s+1)^{-1}\mathrm{E}(X_i^{2s+1})\mathrm{E}(X_i^{2s})^{-1}$. We have chosen this example because it offers a nice heuristic explanation of why the sign of the MLE of θ is asymptotically equivalent to the sign of T_n. Specifically, Taylor-expanding $\exp(-\theta X_i)$, we obtain that (23) is the same as

$$W_i = \theta^s \frac{(-1)^s X_i^s}{s!} + \theta^{s+1} \frac{(-1)^{s+1} X_i^{s+1}}{(s+1)!} + o(\theta^{s+1}) + \varepsilon_i. \tag{25}$$

Letting β_1 denote θ^s, β_2 denote $\mathrm{sgn}(\theta)$, and defining $\tilde{X}_{1i} = (s!)^{-1}(-1)^s X_i^s$ and $\tilde{X}_{2i} = \{(s+1)!\}^{-1}(-1)^{s+1} X_i^{s+1} - \gamma(s!)^{-1}(-1)^s X_i^s$, the model (25) with unrestricted θ is the same as the model

$$W_i = \beta_1(1 - \beta_1^{1/s}\beta_2)\tilde{X}_{1i} + \beta_2\beta_1\beta_1^{1/s}\tilde{X}_{2i} + o(\beta_1^{(s+1)/s}) + \varepsilon_i, \tag{26}$$

where $\beta_1 \geqslant 0$ and $\beta_2 \in \{-1, 1\}$. Because, by construction, $\tilde{X}_{1i}$ and $\tilde{X}_{2i}$ are uncorrelated, the MLEs of $\beta_1(1 - \beta_1^{1/s}\beta_2)$ and of $\beta_2\beta_1\beta_1^{1/s}$ are asymptotically independent. Furthermore, under $\beta_1 = 0$, i.e. under $\theta = 0$, the MLE of $\beta_1(1 - \beta_1^{1/s}\beta_2)$ is asymptotically equivalent to the MLE

$\hat{\beta}_1$ of β_1. Thus, asymptotically, inferences about β_1 in model (26) are equivalent to inferences about β_1 under the linear regression model $W_i = \beta_1 \tilde{X}_{1i} + \varepsilon_i$, with $\beta_1 \geqslant 0$. The MLE of $\beta_2 \beta_1 \beta_1^{1/s}$ in (26) is asymptotically equivalent to the least-squares estimator in the linear regression of W_i on $\tilde{X}_{2i}$. Thus, conditional on $\hat{\beta}_1 > 0$, the MLE of β_2 is asymptotically equivalent to the sign of the MLE of $\beta_2 \beta_1 \beta_1^{1/s}$, i.e. the sign of

$$\sum W_i \tilde{X}_{2i},$$

which under $\beta_1 = 0$ is asymptotically equivalent to the sign of (24).

4. Inferences in multidimensional parametric models

4.1. Introduction

In this Section, we consider the estimation of a $p \times 1$ parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ indexing the law $f(y; \theta)$ when the information matrix at a point θ^* is singular and of rank $p - 1$ and $\log f(y; \theta)$ is differentiable up to a specific order. In what follows $S_j(\theta)$ denotes the score with respect to θ_j, $\partial \log f(Y; \theta)/\partial \theta_j$, $1 \leqslant j \leqslant p$, and S_j denotes $S_j(\theta^*)$. The rank of the information matrix at θ^* is $p - 1$ if and only if $p - 1$ elements of the score vector, say the last $p - 1$ scores,

$$S_2, \ldots, S_p \text{ are linearly independent} \tag{27}$$

and the remaining score is equal to a linear combination of them, i.e.

$$S_1 = K(S_2, \ldots, S_p)^{\mathrm{T}} \tag{28}$$

for some $1 \times (p - 1)$ vector of constants K. In this Section we show that when (27) and (28) hold the MLEs of some or all of the components of θ will converge at rates slower than $O_p(n^{-1/2})$. Furthermore, we derive the asymptotic distribution of the MLE of θ^* and of the likelihood ratio test statistic for testing $H_0 : \theta = \theta^*$ versus $H_1 : \theta \neq \theta^*$.

The informal derivation of the asymptotic distribution of the MLE of θ in Section 3 relied heavily on the first and possibly higher-order derivatives of the log-likelihood at θ^* being identically equal to 0. When θ has more than one component, the singularity of the information matrix does not imply the vanishing of any of the scores corresponding to the components of θ, so the derivations in Section 3 cannot be directly extended to the multiparameter problem. Our derivation of the asymptotic distribution of the MLE $\hat{\theta}$ of θ^* and of the likelihood ratio test is carried out in two steps. First, we consider the more tractable special case in which the following two assumptions that resemble the key conditions of the one-dimensional problem are satisfied: (a) the score corresponding to θ_1 is zero at θ^*, i.e. $S_1 = 0$ and $K = 0$ in equation (28); and (b) higher-order partial derivatives of the log-likelihood with respect to θ_1 at θ^* are possibly also zero, but the first non-zero higher-order partial derivative of the log-likelihood with respect to θ_1 evaluated at θ^* is not a linear combination of the scores $S_2, \ldots, S_p$. Analogously to the one-parameter problem, we show that for this case there exists a (positive integer) power of $\theta_1 - \theta^*$ that is estimable at

rate $\sqrt{n}$. When the parity of the power is odd the asymptotic distribution of the likelihood ratio test statistic is chi-squared with p degrees of freedom. When the parity of the power is even, the likelihood ratio statistic behaves asymptotically as that of an experiment based on data $Y_1, \ldots, Y_n$ in which $\theta_1 - \theta^*$ is known to lie in a closed half real line. Specifically, the asymptotic distribution is a mixture of chi-squared distributions with p and $p-1$ degrees of freedom and mixing probabilities equal to $1/2$. Next, for a general model with information matrix of rank $p-1$ at θ^*, we reduce the derivation of the asymptotic distribution of the desired statistics by working with a reparametrization of the model that satisfies (a) and (b).

4.2. Assumptions

We assume that $Y_1, Y_2, \ldots, Y_n$ are n independent copies of a random variable Y with density $f(y; \theta^*)$ with respect to a carrying measure. Let $l(y; \theta)$ denote $\log f(y; \theta)$ and, for any $1 \times p$ vector $r = (r_1, \ldots, r_p)$, let $l^{(r)}(y; \theta)$ denote $\partial^{r.} \log f(y; \theta)/\partial^{r_1}\theta_1 \partial^{r_2}\theta_2 \ldots \partial^{r_p}\theta_p$, where $r. \equiv \sum_{k=1}^{p} r_k$. Write $L_n(\theta)$ and $L_n^{(r)}(\theta)$ for $\sum l(Y_i; \theta)$ and $\sum l^{(r)}(Y_i; \theta)$ respectively, and define $\|\theta\|^2 = \sum_{k=1}^{p} \theta_k^2$. We assume the following regularity conditions:

(A1) θ^* takes its value in a compact subset Θ of $\mathbb{R}^p$ that contains an open neighbourhood $\mathscr{N}$ of θ^*.

(A2) Distinct values of θ in Θ correspond to distinct probability distributions.

(A3) $\mathrm{E}\{\sup_{\theta \in \Theta} |l(Y; \theta)|\} < \infty$.

(A4) With probability 1, the derivative $l^{(r)}(Y; \theta)$ exists for all θ in $\mathscr{N}$ and $r. \leqslant 2s+1$ and satisfies $\mathrm{E}\{\sup_{\theta \in \mathscr{N}} |l^{(r)}(Y; \theta)|\} < \infty$. Furthermore, with probability 1 under θ^*, $f(Y; \theta) > 0$ for all θ in $\mathscr{N}$.

(A5) For $s \leqslant r. \leqslant 2s+1$, $\mathrm{E}[\{l^{(r)}(Y; \theta^*)\}^2] < \infty$.

(A6) When $r. = 2s+1$ there exists $\varepsilon > 0$ and some function $g(Y)$ satisfying $\mathrm{E}\{g(Y)^2\} < \infty$ such that for θ and θ' in $\mathscr{N}$, with probability 1,

$$|L_n^{(r)}(\theta) - L_n^{(r)}(\theta')| \leqslant |\theta - \theta'|^{\varepsilon} \sum g(Y_i). \tag{29}$$

(A7) Conditions (27) and (28) hold with probability 1 for some $1 \times (p-1)$ constant vector K.

We initially require the following additional conditions:
Let $S_1^{(s+j)}$, $j = 0, 1$, denote $\partial^{s+j} l(Y; \theta)/\partial \theta_1^{s+j}|_{\theta^*}$.

(B1) With probability 1, $\partial^j l(Y; \theta)/\partial \theta_1^j|_{\theta^*} = 0$, $1 \leqslant j \leqslant s-1$.

(B2) For all $1 \times (p-1)$ vectors K, $S_1^{(s)} \neq K(S_2, \ldots, S_p)^{\mathrm{T}}$ with positive probability.

(B3) If s is even, then for all $1 \times p$ vectors K', $S_1^{(s+1)} \neq K'(S_1^{(s)}, S_2, \ldots, S_p)^{\mathrm{T}}$ with positive probability.

4.3. Asymptotic results under (A1)–(A7) and (B1)–(B3)

The following theorems state the asymptotic distribution of $\hat{\theta}$ and of the likelihood ratio test statistic $2\{L_n(\hat{\theta}) - L_n(\theta^*)\}$ when $l(y; \theta)$ satisfies conditions (A1)–(A7) and (B1)–(B3). As

in the one-dimensional problem, the asymptotic behaviour depends on the parity of s. In what follows, I denotes the covariance matrix of $(S_1^{(s)}/s!, S_2, \ldots, S_p)$, I^{jk} denotes the (j, k)th entry of I^{-1}, $Z = (Z_1, Z_2, \ldots, Z_p)^{\mathrm{T}}$ denotes a mean-zero normal random vector with variance equal to I^{-1} and B denotes a Bernoulli variable with success probability equal to $1/2$ that is independent of Z.

Theorem 3. *Under (A1)–(A7) and (B1)–(B2), when s is odd: (a) the MLE $\hat{\theta}$ of θ exists when $\theta = \theta^*$, it is unique with a probability tending to 1, and it is a consistent estimator of θ when $\theta = \theta^*$; (b)*

$$\begin{bmatrix} n^{1/(2s)}(\hat{\theta}_1 - \theta_1^*) \\ n^{1/2}(\hat{\theta}_2 - \theta_2^*) \\ \vdots \\ n^{1/2}(\hat{\theta}_p - \theta_p^*) \end{bmatrix} \rightsquigarrow \begin{bmatrix} Z_1^{1/s} \\ Z_2 \\ \vdots \\ Z_p \end{bmatrix};$$

and (c) $2\{L_n(\hat{\theta}) - L_n(\theta^*)\} \rightsquigarrow \chi^2_p$.

Theorem 4. *Under (A1)–(A7) and (B1)–(B3), when s is even: (a) the MLE $\hat{\theta}$ of θ exists when $\theta = \theta^*$, it is unique with a probability tending to 1, and it is a consistent estimator of θ when $\theta = \theta^*$; (b)*

$$\begin{bmatrix} n^{1/(2s)}(\hat{\theta}_1 - \theta_1^*) \\ n^{1/2}(\hat{\theta}_2 - \theta_2^*) \\ \vdots \\ n^{1/2}(\hat{\theta}_p - \theta_p^*) \end{bmatrix} \rightsquigarrow \begin{bmatrix} (-1)^B Z_1^{1/s} \\ Z_2 \\ \vdots \\ Z_p \end{bmatrix} I(Z_1 > 0) + \begin{bmatrix} 0 \\ Z_2 - (I^{21}/I^{11})Z_1 \\ \vdots \\ Z_p - (I^{p1}/I^{11})Z_1 \end{bmatrix} I(Z_1 < 0), \qquad (30)$$

and (c)

$$2\{L_n(\hat{\theta}) - L_n(\theta^*)\} \rightsquigarrow \sum_{j=1}^{p} Z_j^{*2} I(Z_1^* > 0) + \sum_{j=2}^{p} Z_j^{*2} I(Z_1^* < 0),$$

where Z_j^*, $j = 1, 2, \ldots, p$, are independent $N(0, 1)$ random variables. That is, the asymptotic distribution of the likelihood ratio test statistic is a mixture of a χ^2_{p-1} and a χ^2_p random variable, with mixing probabilities equal to $1/2$.

The second term of the limiting random vector in (30) is the MLE of the mean of Z when the mean of the first component Z_1 is known to be 0. The differences $Z_j - (I^{j1}/I^{11})Z_1$ are the residuals $Z_j - \mathrm{E}(Z_j|Z_1)$ from the population (linear) regression of Z_j on Z_1.

In Theorems 3 and 4, the covariance of Z, I^{-1}, is the Bhattacharyya variance bound of order s for unbiased estimators of $[(\theta_1 - \theta_1^*)^s, \theta_2, \ldots, \theta_p]$ evaluated at $\theta = \theta^*$. Bhattacharyya (1946) gives the multivariate generalization of his univariate bound given in Section 3.

In the proof of Theorem 4 (equation (33)) we show that when s is even and for $n^{1/2}[(\theta_1 - \theta_1^*)^s, \theta_2 - \theta_2^*, \ldots, \theta_p - \theta_p^*]$ bounded, the log-likelihood is, up to an $o_p(1)$ term, a function of $|\theta_1 - \theta_1^*|, \theta_2, \ldots, \theta_p$. Thus, the calculation of the statistic that asymptotically determines the sign of the MLE of $\theta - \theta^*$ requires a higher-order expansion of the log-likelihood. This is shown in the Appendix to effectively yield estimation of the sign of $\theta_1 - \theta_1^*$ from the sign of the product of the MLE of $\theta - \theta^*$ and the sum of residuals of the population least-squares regression of the vector

$$[l^{[s+1,0,\ldots,0]}(Y_i, \theta^*), l^{(1,1,0,\ldots,0)}(Y_i, \theta^*), \ldots, l^{(1,0,\ldots,0,1)}(Y_i, \theta^*)]$$

on the vector

$$[l^{[s,0,\ldots,0]}(Y_i, \theta^*), l^{(0,1,0,\ldots,0)}(Y_i, \theta^*), \ldots, l^{(0,0,\ldots,0,1)}(Y_i, \theta^*)], \qquad i = 1, \ldots, n,$$

(equation (44), where $l^{[s+j,0,\ldots,0]}(Y_i, \theta^*) = l^{(s+j,0,\ldots,0)}(Y_i, \theta^*)/(s+j)!$. If interest is focused on θ_1, the other components of the vector θ being regarded as nuisance parameters, an analysis of the local behaviour of inferences near $\theta_1 = \theta_1^*$ similar to that carried out in Section 3 would reveal several possibilities as in the one-parameter problem. In particular, it would indicate the possibility of profile likelihood confidence regions for θ_1 being comprised of two disjoint intervals located at each side of θ_1^* when the data are generated under $\theta_1 = an^{-b}$ and $1/(2s+2) \leqslant b < 1/(2s)$.

The results of Theorem 4 when s is even are strongly connected with the results of Geyer (1994) on maximum likelihood estimation subject to a boundary constraint. See also the related results under more stringent regularity conditions of Moran (1971) on the distribution of the MLE, and of Chant (1974) and Self and Liang (1987) on the distribution of the likelihood ratio test statistic, the latter drawing from the earlier work of Chernoff (1954) on the distribution of the likelihood ratio test statistic of a composite null hypothesis when the true parameter is in the interior of the parameter space but on the boundary of the parameter sets defining the null and alternative hypothesis. The essence of the connection is that the asymptotic distribution of the MLE of $[(\theta_1 - \theta_1^*)^s, \theta_2, \ldots, \theta_p]$ at $\theta = \theta^*$ in the submodel in which θ_1 is known to satisfy $\theta_1 \geqslant \theta_1^*$ can be obtained from the results of Geyer (1994) after reparametrization. This distribution agrees with the asymptotic distribution of the MLE of $[(\theta_1 - \theta_1^*)^s, \theta_2, \ldots, \theta_p]$ given in Theorem 4. Thus inference about $[|\theta_1 - \theta_1^*|, \theta_2, \ldots, \theta_p]$ is unaffected by the constraint $\theta_1 \geqslant \theta_1^*$.

4.4. Asymptotic results under (A1)–(A7) when (B1)–(B3) do not hold

We now derive the asymptotic distributions of the MLE and of the likelihood ratio test statistic when conditions (A1)–(A7) hold but conditions (B1)–(B3) are not true. The fundamental idea behind our derivation is the iterative reparametrization of the model until conditions (B1)–(B3) are satisfied. Specifically, when equation (28) holds with $K \neq 0$, we start with a reparametrization $\psi = \psi(\theta)$ such that: (i) the scores corresponding to $\psi_2, \ldots, \psi_p$ evaluated at $\psi^* = \psi(\theta^*)$ are equal to the scores corresponding to $\theta_2, \ldots, \theta_p$ in the originally parametrized model evaluated at θ^*; and (ii) the score corresponding to ψ_1 evaluated at ψ^* is orthogonal to, i.e. uncorrelated with, the vector of scores corresponding to $\psi_2, \ldots, \psi_p$

evaluated at ψ^*. Conditions (i) and (ii) imply that the score for ψ_1 evaluated at ψ^* is simultaneously orthogonal and linearly dependent with the vector of scores for $\psi_2, \ldots, \psi_p$ evaluated at ψ^*, which in turn implies that the score for ψ_1 evaluated at ψ^* is equal to 0. The reparametrization $\psi = \theta + [0, K^{\mathrm{T}}]^{\mathrm{T}}(\theta_1 - \theta_1^*)$ satisfies conditions (i) and (ii) (Cox and Reid 1987b). Notice that under this reparametrization $\psi_1 = \theta_1$ and $\psi^* = \theta^*$. Furthermore, the constant K satisfying condition (28) is equal to the population least-squares projection constant, i.e. $K = \mathrm{E}(S_1\Gamma)\mathrm{E}(\Gamma\Gamma^{\mathrm{T}})^{-1}$, where $\Gamma = (S_2, \ldots, S_p)^{\mathrm{T}}$. In the model parametrized by ψ the score for ψ_1 evaluated at ψ^* is equal to 0. To check if conditions (B1)–(B2) are satisfied in this model with $s = 2$, we need to evaluate the second partial derivative with respect to ψ_1 of the reparametrized log-likelihood evaluated at ψ^*. If it is neither equal to zero nor a linear combination of the scores for the remaining parameters, then the reparametrized model satisfies conditions (B1) and (B2) with $s = 2$. Otherwise, we set K_2 equal to the coefficients of the linear combination or $K_2 = 0$ if this derivative is zero. That is, we set K_2 equal to the population least-squares projection constant, $\mathrm{E}(\tilde{S}_1^{(2)}\Gamma)\mathrm{E}(\Gamma\Gamma^{\mathrm{T}})^{-1}$, where $\tilde{S}_1^{(2)}$ is the second partial derivative of the reparametrized log-likelihood with respect to ψ_1 evaluated at ψ^*. Next, we consider the new reparametrization $\psi = \theta + [0, K]^{\mathrm{T}}(\theta_1 - \theta_1^*) + [0, 1/2K_2]^{\mathrm{T}}(\theta_1 - \theta_1^*)^2$. The newly reparametrized model satisfies $\psi_1 = \theta_1$ and $\psi = \theta^*$ when $\theta = \theta^*$. Under the new parametrization: (a) the scores for $\psi_2, \ldots, \psi_p$ evaluated at $\psi^* = \psi(\theta^*)$ remain unchanged; and (b) the second partial derivative of the log-likelihood with respect to ψ_1 evaluated at ψ^* and the vector of scores for $\psi_2, \ldots, \psi_p$ evaluated at ψ^* are orthogonal. Thus, in particular, in the newly reparametrized model the first and second partial derivatives of the log-likelihood with respect to ψ_1 evaluated at ψ^* are equal to zero and the scores for the remaining parameters are equal to the scores for $\theta_2, \ldots, \theta_p$ in the originally parametrized model. For the newly parametrized model, we now check whether the third partial derivative of the log-likelihood with respect to ψ_1 is neither zero nor a linear combination of the scores for the remaining parameters. If that is the case, the iterative reparametrization stops and the reparametrized model satisfies conditions (B1) and (B2) with $s = 3$. Otherwise, the process of reparametrization continues until condition (B2) is satisfied. If s is even, we now need to check further that condition (B3) is satisfied by the reparametrized model. If this condition fails, further reparametrization is needed. However, in this paper we consider only models in which condition (B3) is satisfied if condition (B2) holds.

We will henceforth assume that the following conditions hold for some positive integer s.

(C1) Set $K_0 = 0$ and A_0 equal to the $p \times 1$ null vector. With probability 1, there exist $1 \times (p-1)$ (possibly null) vectors $K_1, K_2, \ldots, K_{s-1}$ defined iteratively for $1 \leqslant j \leqslant s-1$ by

$$\left.\frac{\partial^j l\left\{Y;\, \theta - \sum_{l=0}^{j-1} A_l(\theta_1 - \theta_1^*)^l\right\}}{\partial \theta_1^j}\right|_{\theta^*} = K_j(S_2, S_3, \ldots, S_p)^{\mathrm{T}}, \tag{31}$$

where A_j denotes the $p \times 1$ vector $[0, (j!)^{-1}K_j]^{\mathrm{T}}$.

(C2) For $j = 0, 1,$ define

$$\tilde{S}_1^{(s+j)} \equiv \left. \frac{\partial^{s+j} l\left\{ Y;\, \theta - \sum_{l=0}^{s-1} A_l(\theta_1 - \theta_1^*)^l \right\}}{\partial \theta_1^{s+j}} \right|_{\theta^*}.$$

Then with probability greater than 0, $\tilde{S}_1^{(s)}$ is neither zero nor a linear combination of $S_2, \ldots, S_p$ and, if s is even, $\tilde{S}_1^{(s+1)}$ is neither zero nor a linear combination of $\tilde{S}_1^{(s)}, S_2, \ldots, S_p$.

Note that K_1 defined in (C1) exists and is equal to K defined in (28). Furthermore, K_j is the population least-squares projection constant of the random variable on the left-hand side of (31) on the vector $(S_2, S_3, \ldots, S_p)$.

Let I denote the covariance of $(\tilde{S}_1^{(s)}/s!, S_2, \ldots, S_p)$, let I^{jk} denote the (j, k)th element of I^{-1}, let $Z = (Z_1, Z_2, \ldots, Z_p)^{\mathrm{T}}$ denote a mean-zero normal random vector with covariance I^{-1} and let B be a Bernoulli random variable with success probability $1/2$ independent of Z.

Theorem 5. *Under (A1)–(A7), (C1) and (C2): (a) the MLE $\hat{\theta}$ of θ exists when $\theta = \theta^*$, it is unique with a probability tending to 1, and it is a consistent estimator of θ when $\theta = \theta^*$; (b) when s is odd, we have*

(i)

$$\begin{bmatrix} n^{1/(2s)}(\hat{\theta}_1 - \theta_1^*) \\ n^{1/2}\{(\hat{\theta}_2 - \theta_2^*) + \sum_{j=0}^{s-1} (j!)^{-1} K_{j1}(\hat{\theta}_1 - \theta_1^*)^j\} \\ \vdots \\ n^{1/2}\{(\hat{\theta}_p - \theta_p^*) + \sum_{j=0}^{s-1} (j!)^{-1} K_{j(p-1)}(\hat{\theta}_1 - \theta_1^*)^j\} \end{bmatrix} \rightsquigarrow \begin{bmatrix} Z_1^{1/s} \\ Z_2 \\ \vdots \\ Z_p \end{bmatrix},$$

where K_{jl} is the lth element of K_j, $0 \leqslant j \leqslant s-1$, $1 \leqslant l \leqslant p-1$,

(ii) $2\{L_n(\hat{\theta}) - L_n(\theta^*)\} \rightsquigarrow \chi_p^2$;

and (c) when s is even, we have

(i)

$$\begin{bmatrix} n^{1/(2s)}(\hat{\theta}_1 - \theta_1^*) \\ n^{1/2}\{(\hat{\theta}_2 - \theta_2^*) + \sum_{j=0}^{s-1} (j!)^{-1} K_{j1}(\hat{\theta}_1 - \theta_1^*)^j\} \\ \vdots \\ n^{1/2}\{(\hat{\theta}_p - \theta_p^*) + \sum_{j=0}^{s-1} (j!)^{-1} K_{j(p-1)}(\hat{\theta}_1 - \theta_1^*)^j\} \end{bmatrix} \rightsquigarrow \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_p \end{bmatrix},$$

where

$$\begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_p \end{bmatrix} = \begin{bmatrix} (-1)^B Z_1^{1/s} \\ Z_2 \\ \vdots \\ Z_p \end{bmatrix} I(Z_1 > 0) + \begin{bmatrix} 0 \\ Z_2 - (I^{21}/I^{11})Z_1 \\ \vdots \\ Z_p - (I^{p1}/I^{11})Z_1 \end{bmatrix} I(Z_1 < 0),$$

(ii) $2\{L_n(\hat{\theta}) - L_n(\theta^*)\} \rightsquigarrow \sum_{j=1}^{p} Z_j^{*2} I(Z_1^* > 0) + \sum_{j=2}^{p} Z_j^{*2} I(Z_1^* < 0)$, *where* Z_j^*, $j = 1, 2, \ldots, p$, *are independent* $N(0, 1)$ *random variables.*

Notice that when conditions (B1)–(B3) hold, $\tilde{S}_1^{(s+j)}$ is equal to $S_1^{(s+j)}$, $j = 0, 1$, because the vectors K_j, $j = 1, \ldots, s-1$, defined in (31) are all equal to 0. Thus, Theorems 3 and 4 are special cases of Theorem 5. We nevertheless stated these theorems separately because our proof of Theorem 5 builds up from the asymptotic results under the more stringent conditions (B1)–(B3).

In proving Theorems 3 and 4 (equation (39)) we show that the difference between the likelihood ratio test statistic and its limiting random variable is of order $O_p(n^{-1/(2s)})$. The difference between the respective cumulative distribution functions is of order $O(n^{-1/(2s)})$ when s is even but it is of order $O(n^{-1/s})$ when s is odd (Cox and Reid 1987a).

4.5. Example

We now apply the iterative reparametrization and the results of Theorem 5 to derive the asymptotic distribution of the estimators of the example of Section 2. Suppose, first, that σ is known and equal to σ^*. Let $\theta = (\alpha_1, \beta, \alpha_0)^{\mathrm{T}}$, and $\theta^* = (\alpha_1^*, \beta^*, \alpha_0^*)^{\mathrm{T}}$, where $\alpha_1^* = 0$, α_0^* is a fixed and arbitrary value, and $\beta^* = 0$ without loss of generality. The individual contribution to the derivative of the log-likelihood evaluated at $\theta = \theta^*$, (S_1, S_2, S_3), is given by

$$[RH'(\alpha_0^*)\sigma^{*-1}Y,\ R\sigma^{*-2}Y,\ \{R - (1-R)A_0\}H'(\alpha_0^*)],$$

where $A_0 = \mathrm{e}^{H(\alpha_0^*)}\{1 - \mathrm{e}^{H(\alpha_0^*)}\}^{-1}$. Since $S_1 = K_{11}\, S_2$, where $K_{11} = \sigma^* H'(\alpha_0^*)$ and S_2 and S_3 are linearly independent, we consider the reparametrization $(\alpha_1, \beta, \alpha_0) \rightarrow (\alpha_1,\ \beta + K_{11}\alpha_1, \alpha_0)$. The second derivative of the log-likelihood with respect to α_1 evaluated at $\theta = \theta^*$ in the reparametrized model is equal to

$$\tilde{S}_1^{(2)} = R[\{H'(\alpha_0^*)\}^2 + H''(\alpha_0^*)\sigma^{*^{-2}}Y^2] - (1-R)[\{H'(\alpha_0^*)\}^2 + H''(\alpha_0^*)]A_0. \qquad (32)$$

Since $\tilde{S}_1^{(2)}$ is a function of Y^2, it cannot be a linear combination of S_2 and S_3. Thus, the iterative reparametrization stops and by Theorem 5,

$$[n^{1/4}\hat{\alpha}_1,\ n^{1/2}(\hat{\beta} + K_{11}\hat{\alpha}_1),\ n^{1/2}(\hat{\alpha}_0 - \alpha_0^*)]$$

converges under $\theta = \theta^*$ to the random vector $W = (W_1, W_2, W_3)$ given in part (c)(i) of that

Theorem with I equal to the covariance of $(\tilde{S}_1^{(2)}/2, S_2, S_3)$. The distribution of W coincides with the asymptotic distribution of the estimators found in Section 2 when σ is known.

Suppose now that σ is unknown. With θ and θ^* redefined as $\theta = (\alpha_1, \beta, \alpha_0, \sigma)^{\mathrm{T}}$ and $\theta^* = (\alpha_1^*, \beta^*, \alpha_0^*, \sigma^*)^{\mathrm{T}}$, the derivative of the log-likelihood with respect to σ evaluated at $\theta = \theta^*$, S_4, is equal to $R\sigma^{*^{-3}}(Y^2 - \sigma^{*^2})$. Since S_4 is linearly independent of S_1, S_2 and S_3, we consider the reparametrization $(\alpha_1, \beta, \alpha_0, \sigma) \to (\alpha_1, \beta + K_{11}\alpha_1, \alpha_0, \sigma)$. In the reparametrized model the individual contribution to the second derivative of the log-likelihood with respect to α_1 evaluated at $\theta = \theta^*$ remains equal to (32). However, now $\tilde{S}_1^{(2)}$ can be written as $K_{22}S_3 + K_{23}S_4$, where $K_{22} = \{H'(\alpha_0^*)\}^{-1}[\{H'(\alpha_0^*)\}^2 + H''(\alpha_0^*)]$ and $K_{23} = \sigma^* H''(\alpha_0^*)$. Thus, we consider the new reparametrization

$$(\alpha_1, \beta, \alpha_0, \sigma) \to (\alpha_1, \beta + K_{11}\alpha_1, \alpha_0 + 2^{-1}K_{22}\alpha_1^2, \sigma + 2^{-1}K_{23}\alpha_1^2).$$

In the newly reparametrized model, the third partial derivative of the log-likelihood with respect to α_1 evaluated at $\theta = \theta^*$ is equal to

$$\tilde{S}_1^{(3)} = -3\{H'(\alpha_0^*)\}^{-1}\{H''(\alpha_0^*)\}^2\sigma^{*-1}Y + H'''(\alpha_0^*)\sigma^{*-3}Y^3.$$

This expression is a cubic polynomial in Y and therefore cannot be a linear combination of the scores S_2, S_3 and S_4. The iterative reparametrization therefore stops and by Theorem 5(b)(i),

$$[n^{1/6}\hat{\alpha}_1, n^{1/2}(\hat{\beta} + K_{11}\hat{\alpha}_1), n^{1/2}(\hat{\alpha}_0 - \alpha_0^* + 2^{-1}K_{22}\hat{\alpha}_1^2), n^{1/2}(\hat{\sigma} - \sigma^* + 2^{-1}K_{23}\hat{\alpha}_1^2)]$$

converges under $\theta = \theta^*$ to a normal random vector with mean 0 and covariance matrix I equal to the covariance of $(\tilde{S}_1^{(3)}/6, S_2, S_3, S_4)$, which agrees with the results derived in Section 2.

5. Asymptotic properties of the second and higher-order derivatives of the log-likelihood

The results in Sections 3 and 4 rely heavily on the asymptotic distribution of the second and higher-order derivatives of the log-likelihood. These distributions are derived here as a consequence of a general result stated in the following lemma on the properties of the high-order derivatives of the logarithm of a function whose first $s - 1$ derivatives vanish at a point. This lemma, implicit from the exlog relations described in Barndorff-Nielsen and Cox (1989, pp. 140–142), is shown in the Appendix to follow immediately from Faà di Bruno's (1859, p. 3) formula on the derivatives of a composition of two functions. Define the $1 \times p$ vectors $r_{(j)} \equiv (j, 0, 0, \ldots, 0)$, $\tilde{R}_{(j)} \equiv (j, 1, 0, \ldots, 0)$, $0 \leqslant j \leqslant 2s + 1$, and $a \equiv (1, 2, 0, \ldots, 0)$.

Lemma 1. *Let* $h : \Omega \to R$, *where* Ω *is an open set of* $\mathbb{R}^p$. *Suppose* $\partial^{2s+1}h(u)/\partial u^{r_1}\partial u^{r_2}\ldots\partial u^{r_p}$, $\sum r_i = 2s + 1$, *exist at an interior point* $u^* = (u_1^*, u_2^*, \ldots, u_p^*)$ *of* Ω. *Define* $G(u) = \log h(u)$, *and for* $r = (r_1, r_2, \ldots, r_p)$ *and* $r. = \sum r_i$ *let*

$$h_r^* = \frac{\{\partial^{r\cdot} h(u)/\partial u_1^{r_1}\partial u_2^{r_2}\ \ldots\ \partial u_p^{r_p}\}|_{u^*}}{h(u^*)},$$

and

$$G^{(r)}(u) = \partial^{r\cdot} G(u)/\partial u_1^{r_1}\partial u_2^{r_2}\ \ldots\ \partial u_p^{r_p}.$$

Then:

(i) *for* $1 \leqslant j \leqslant s-1$, $h^*_{r_{(j)}} = 0$ *if and only if* $G^{(r_{(j)})}(u^*) = 0$.
Furthermore, if $G^{(r_{(1)})}(u^*) = \ldots = G^{(r_{(s-1)})}(u^*) = 0$ *then:*

(ii) *for* $s \leqslant j \leqslant 2s-1$, $G^{(r_{(j)})}(u^*) = h^*_{r_{(j)}}$;

(iii) $G^{(r_{(2s)})}(u^*) = h^*_{r_{(2s)}} - \{(2s)!/2(s!)^2\}\{G^{(r_{(s)})}(u^*)\}^2$;

(iv) $G^{(r_{(2s+1)})}(u^*) = h^*_{r_{(2s+1)}} - \binom{2s+1}{s} G^{(r_{(s)})}(u^*) G^{(r_{(s+1)})}(u^*)$;

(v) *for* $1 \leqslant j \leqslant s-1$, $G^{(\tilde{R}_{(j)})}(u^*) = h^*_{\tilde{R}_{(j)}}$;

(vi) $G^{(\tilde{R}_{(s)})}(u^*) = h_{\tilde{R}_{(s)}} - G^{(r_{(s)})}(u^*) G^{(\tilde{R}_{(0)})}(u^*)$;

(vii) $G^{(\tilde{R}_{(s+1)})}(u^*) = h^*_{\tilde{R}_{(s+1)}} - (s+1) G^{(r_{(s)})}(u^*) G^{(r_{(1)})}(u^*) - G^{(r_{(s+1)})}(u^*) G^{(\tilde{R}_{(0)})}(u^*)$;

(viii) $G^{(a)}(u^*) = h^*_a - 2G^{(\tilde{R}_{(1)})}(u^*) G^{(\tilde{R}_{(0)})}(u^*)$;

(ix) *parts* (v), (vi) *and* (vii) *are also true if* $\tilde{R}_{(j)}$ *and* a *are defined as before but* $\tilde{R}_{(j)}$ *has 0 in its second entry and 1 in its* k*th entry, and* a *has 0 in its second entry and 2 in its* k*th entry, and* k *is any fixed index between 3 and* p.

The following corollary of Lemma 1 follows immediately from the central limit theorem under the regularity conditions (A1), (A4), (A5) and (B1) of Section 4.2, since under conditions (A1), (A4) and (A5), $f^{(r)}(Y;\theta^*)/f(Y;\theta^*)$ has zero mean and $l^{(r)}(Y;\theta^*)$ has finite variance for all r with $r. \leqslant 2s+1$, and under condition (B1) the derivatives $l^{(r)}(Y;\theta^*)$ can be obtained from Theorem 5. The corollary states the asymptotic behaviour of the derivatives of the log-likelihood. For any $r = (r_1, r_2, \ldots, r_p)$ let $L_n^{(r)}(\theta)$ denote $\sum l^{(r)}(Y_i;\theta)$ and write $l^{[r]}(Y;\theta^*)$ for $f^{(r)}(Y;\theta^*)/\{m! f(Y;\theta^*)\}$, where $m = \max\{r_2, \ldots, r_p\}$.

Corollary 1. *Define*

$$I_{11} \equiv \mathrm{E}[\{l^{[r_{(s)}]}(Y;\theta^*)\}^2], \qquad I_{22} \equiv \mathrm{E}[\{l^{[\tilde{R}_{(0)}]}(Y;\theta^*)\}^2],$$

$$I_{21} \equiv I_{12} \equiv \mathrm{E}\{l^{[r_{(s)}]}(Y;\theta^*) l^{[\tilde{R}_{(0)}]}(Y;\theta^*)\},$$

$$C_{11} \equiv \mathrm{E}\{l^{[r_{(s)}]}(Y;\theta^*) l^{[r_{(s+1)}]}(Y;\theta^*)\}, \qquad C_{22} \equiv \mathrm{E}\{l^{[\tilde{R}_{(0)}]}(Y;\theta^*) l^{[\tilde{R}_{(1)}]}(Y;\theta^*)\},$$

$$C_{12} \equiv \mathrm{E}\{l^{[r_{(s)}]}(Y;\theta^*) l^{[\tilde{R}_{(1)}]}(Y;\theta^*)\}, \qquad C_{21} \equiv \mathrm{E}\{l^{[\tilde{R}_{(0)}]}(Y;\theta^*) l^{[r_{(s+1)}]}(Y;\theta^*)\}.$$

Then, under assumptions (A1), (A4), (A5) and (B1) of Section 4.2:

(a) *For* $s \leqslant j \leqslant 2s-1$,

$$n^{-1/2} L_n^{(r_{(j)})}(\theta^*) = j! n^{-1/2} \sum_i^n l^{[r_{(j)}]}(Y_i;\theta^*)$$
$$= Z_{r_{(j)}} + o_p(1),$$

where $Z_{r_{(j)}}$ *is a mean-zero normal random variable. The variance of* $Z_{r_{(s)}}$ *is equal to* $(s!)^2 I_{11}$.

(b) For $1 \leqslant j \leqslant s-1$,

$$n^{-1/2} L_n^{(\tilde{R}_{(j)})}(\theta^*) = (j!) n^{-1/2} \sum_i^n l^{[\tilde{R}_{(j)}]}(Y_i;\, \theta^*)$$
$$= Z_{\tilde{R}_{(j)}} + o_p(1),$$

where $Z_{\tilde{R}_{(j)}}$ *is a mean-zero normal random variable.*

(c) $n^{-1} L_n^{(r_{(2s)})}(\theta^*) = -\{(2s)!/2\} I_{11} + O_p(n^{-1/2})$.

(d) $n^{-1} L_n^{(r_{(2s+1)})}(\theta^*) = -\{(2s+1)!\} C_{11} + O_p(n^{-1/2})$.

(e) $n^{-1} L_n^{(\tilde{R}_{(s)})}(\theta^*) = -s! I_{12} + O_p(n^{-1/2})$.

(f) $n^{-1} L_n^{(\tilde{R}_{(s+1)})}(\theta^*) = -(s+1)!\{C_{12} + C_{21}\} + O_p(n^{-1/2})$.

(g) $n^{-1} L_n^{(a)}(\theta^*) = -2C_{22} + O_p(n^{-1/2})$.

6. Discussion

There are a number of directions in which the results of the present paper need to be extended. The information matrix at $\theta = \theta^*$ may be of rank $p-q$, where $1 < q \leqslant p$. For example, in the model of Section 2, the rank is $p-2$ if $H'(\alpha_0) = 0$. Then the structure of the maximum likelihood estimate is similar to but more complex than that derived above for $q = 1$, and the law of the likelihood ratio test statistic is a mixture of chi-squared distributions.

Even when conditions (B1) and (B2) of Section 4.2 hold, it is possible for the $(s+1)$th and sth derivatives of the log-likelihood with respect to θ_1 to be linearly dependent. Then the calculation of the statistic that asymptotically determines the sign of $\hat{\theta}_1 - \theta_1^*$ requires examination of higher-order terms of the log-likelihood expansion and, in some cases, additional reparametrization.

If the parameter θ is partitioned into a parameter ξ of interest and a nuisance parameter ϕ, then inference about ξ independent of ϕ may raise special problems near $\theta = \theta^*$, especially if there is serious ambiguity over the sign of certain components in ϕ. For example, in the special case discussed in Section 2, with the mean β as the parameter of interest, the adjustment to the sample mean has a sign determined by the sign of α_1; it can then happen that the magnitude but not the direction of the adjustment is fairly well determined. Then the confidence set would be a pair of intervals. An interesting consequence is that in such cases intervals for β obtained from the profile likelihood sometimes have the wrong structure. This and more complex issues will not be discussed in the present paper.

Appendix

Proofs of Theorems 1–4

Theorems 1 and 2 are special cases of Theorems 3 and 4 when $p = 1$. Therefore we only

need to prove Theorems 3 and 4. For simplicity of notation we prove the result for $p = 2$ and state at the end of the proof the additional steps required to extend the result for an arbitrary p. Let $\tilde{\theta} \equiv (\tilde{\theta}_1, \tilde{\theta}_2^{\mathrm{T}})$ be a sequence satisfying $[(\tilde{\theta}_1 - \theta_1^*)^s,\ (\tilde{\theta}_2 - \theta_2^*)^{\mathrm{T}}] = O_p(n^{-1/2})$. When $p = 2$, θ_2 is a scalar. We nevertheless write the transpose of θ_2, indicated by the superscript T, and use scalar transposition occasionally throughout the proof, to facilitate the extension later to an arbitrary p. Throughout, if v is a vector of dimension p, we write $v = O_p(n^{-\alpha})$ and $v = o_p(n^{-\alpha})$ to indicate that all the elements of v are $O_p(n^{-\alpha})$ and $o_p(n^{-\alpha})$, respectively. Let $(\tilde{\omega}_1, \tilde{\omega}_2^{\mathrm{T}}) \equiv [n^{1/(2s)}(\tilde{\theta}_1 - \theta_1^*),\ \ n^{1/2}(\tilde{\theta}_2 - \theta_2^*)^{\mathrm{T}}]$ and $L_n^{(j_1,j_2)}(\theta) \equiv \partial^{j_1+j_2} L_n(\theta)/\partial\theta_1^{j_1}\partial\theta_2^{j_2}$. Denote $L_n^{(j_1,j_2)}(\theta^*)$ by $L_n^{(j_1,j_2)}$. By assumption (B1) $L_n^{(j,0)} = 0,\ 1 \leqslant j \leqslant s-1$, so, forming the Taylor expansion of $L_n(\tilde{\theta})$, around θ^* we obtain

$$
\begin{aligned}
L_n(\tilde{\theta}) = L_n(\theta^*) &+ \tilde{\omega}_1^s \left[\left\{ n^{-1/2} \frac{L_n^{(s,0)}}{s!} \right\} \right. \\
&+ n^{-1/(2s)} \left\{ n^{-1/2} \frac{L_n^{(s+1,0)}}{(s+1)!} \tilde{\omega}_1 \right\} + n^{-1/(2s)} \left\{ \sum_{j_1=2}^{s-1} n^{-1/2} \frac{L_n^{(s+j_1,0)}}{(s+j_1)!} n^{(1-j_1)/(2s)} \tilde{\omega}_1^{j_1} \right\} \\
&\left. + \left\{ n^{-1} \frac{L_n^{(2s,0)}}{(2s)!} \tilde{\omega}_1^s \right\} + n^{-1/(2s)} \left[\left\{ n^{-1} \frac{L_n^{(2s+1,0)}}{(2s+1)!} \tilde{\omega}_1^{s+1} \right\} + \left\{ n^{-1} \frac{\delta_n^{(2s+1,0)}}{(2s+1)!} \tilde{\omega}_1^{s+1} \right\} \right] \right] \\
&+ \tilde{\omega}_2^{\mathrm{T}} \left[\left\{ n^{-1/2} L_n^{(0,1)} \right\} + n^{-1/(2s)} \left[\left\{ n^{-1/2} L_n^{(1,1)} \tilde{\omega}_1 \right\} + \left\{ \sum_{j_1=2}^{s-1} n^{-1/2} \frac{L_n^{(j_1,1)}}{j_1!} n^{(1-j_1)/(2s)} \tilde{\omega}_1^{j_1} \right\} \right] \right. \\
&+ \left\{ n^{-1} \frac{L_n^{(s,1)}}{s!} \tilde{\omega}_1^s \right\} + n^{-1/(2s)} \left[\left\{ n^{-1} \frac{L_n^{(s+1,1)}}{(s+1)!} \tilde{\omega}_1^{s+1} \right\} + \left\{ \sum_{j_1=s+2}^{2s} n^{-1} \frac{L_n^{(j_1,1)}}{j_1!} n^{(1-j_1)/(2s)} \tilde{\omega}_1^{j_1} \right\} \right. \\
&\left. \left. + \left\{ n^{-1} \frac{\delta_n^{(2s,1)}}{(2s)!} n^{(1-s)/(2s)} \tilde{\omega}_1^{2s} \right\} \right] + \left\{ n^{-1} \frac{L_n^{(0,2)}}{2} \tilde{\omega}_2 \right\} + n^{-1/(2s)} \left[\left\{ n^{-1} \frac{L_n^{(1,2)}}{2} \tilde{\omega}_1 \tilde{\omega}_2 \right\} \right. \\
&+ \left. \left\{ \sum_{j_1=2}^{2s-1} n^{-1} \frac{L_n^{(j_1,2)}}{(2+j_1)!} n^{(1-j_1)/(2s)} \tilde{\omega}_1^{j_1} \tilde{\omega}_2 \binom{2+j_1}{j_1} \right\} \right. \\
&\left. + \left\{ n^{-1} \frac{\delta_n^{(2s-1,2)}}{(2s+1)!} n^{(2-2s)/(2s)} \tilde{\omega}_1^{2s-1} \tilde{\omega}_2 \binom{2s+1}{2s-1} \right\} \right]
\end{aligned}
$$

$$+\left\{\sum_{k=3}^{2s+1}\sum_{\substack{j_1+j_2=k\\ j_2\geqslant 3, j_1\geqslant 0}} n^{-1}\frac{L_n^{(j_1,j_2)}}{k!}n^{\{(1-j_1)+(2-j_2)s\}/(2s)}\tilde{\omega}_1^{j_1}\tilde{\omega}_2^{j_2-1}\binom{k}{j_1}\right\}$$

$$+\left\{\sum_{\substack{j_1+j_2=2s+1\\ j_2\geqslant 3, j_1\geqslant 0}} n^{-1}\frac{\delta_n^{(j_1,j_2)}}{(2s+1)}n^{\{(1-j_1)+(2-j_2)s\}/(2s)}\tilde{\omega}_1^{j_1}\tilde{\omega}_2^{j_2-1}\binom{2s+1}{j_1}\right\}\Bigg]\Bigg]$$

$$\begin{aligned}
&= L_n(\theta^*)+\tilde{\omega}_1^s(A_{1n}+n^{-1/(2s)}A_{2n}+n^{-1/(2s)}A_{3n}+A_{4n}+n^{-1/(2s)}(A_{5n}+A_{6n}))\\
&\quad+\tilde{\omega}_2^{\mathrm{T}}\{A_{7n}+n^{-1/(2s)}(A_{8n}+A_{9n})+A_{10n}+n^{-1/(2s)}(A_{11n}+A_{12n}+A_{13n})\\
&\quad+A_{14n}+n^{-1/(2s)}(A_{15n}+A_{16n}+A_{17n}+A_{18n}+A_{19n})\},
\end{aligned}$$

where for $j_1+j_2=2s+1$, $\delta_n^{(j_1,j_2)}=L_n^{(j_1,j_2)}(\bar{\theta})-L_n^{(j_1,j_2)}(\theta^*)$, for some $\bar{\theta}$ satisfying $\|\bar{\theta}-\theta^*\|\leqslant\|\theta-\theta^*\|$ and the terms A_{jn}, $1\leqslant j\leqslant 19$ correspond to the terms in braces in the order they appear.

By Corollary 1(a)–(b) of Section 5 and by $\tilde{\omega}_1=O_p(1)$ and $\tilde{\omega}_2=O_p(1)$, we have that $A_{3n}=o_p(1)$ and $A_{9n}=o_p(1)$. By assumption (A5) and the weak law of large numbers, $n^{-1}L_n^{(j_1,j_2)}=O_p(1)$ for $j_1+j_2\leqslant 2s+1$ and hence by $\tilde{\omega}_1=O_p(1)$ and $\tilde{\omega}_2=O_p(1)$, $A_{12n}=o_p(1)$, $A_{16n}=o_p(1)$ and $A_{18n}=o_p(1)$. By assumption (A6), and by $\tilde{\omega}_1=O_p(1)$ and $\tilde{\omega}_2=O_p(1)$, $A_{6n}=o_p(1)$, $A_{13n}=o_p(1)$, $A_{17n}=o_p(1)$, and $A_{19n}=o_p(1)$. By Corollary 1(c)–(g),

$$\begin{aligned}
A_{4n}&=\{-I_{11}/2+o_p(n^{-1/(2s)})\}\tilde{\omega}_1^s,\quad A_{5n}=\{-C_{11}+o_p(n^{-1/(2s)})\}\tilde{\omega}_1^{s+1},\\
A_{10n}&=\{-I_{21}+o_p(n^{-1/(2s)})\}\tilde{\omega}_1^s,\quad A_{11n}=\{-C_{12}^{\mathrm{T}}-C_{21}+o_p(n^{-1/(2s)})\}\tilde{\omega}_1^{s+1},\\
A_{14n}&=\{-I_{22}/2+o_p(n^{-1/(2s)})\}\tilde{\omega}_2,\quad A_{15n}=\{-C_{22}+o_p(n^{-1/(2s)})\}\tilde{\omega}_1\tilde{\omega}_2,
\end{aligned}$$

where I_{jk} and C_{jk}, $j,k=1,2$, are defined in Section 5. Thus, regrouping terms, we obtain

$$L_n(\tilde{\theta})=L_n(\theta^*)+G_n^*(\tilde{\omega}_1^s,\tilde{\omega}_2)+R_n^*(\tilde{\omega}_1,\tilde{\omega}_2), \tag{33}$$

where

$$\begin{aligned}
G_n^*(\tilde{\omega}_1^s,\tilde{\omega}_2)&=(\tilde{\omega}_1^s,\tilde{\omega}_2^{\mathrm{T}})[\{n^{-1/2}L_n^{(s,0)}/s!,(n^{-1/2}L_n^{(0,1)})^{\mathrm{T}}\}^{\mathrm{T}}-\tfrac{1}{2}I(\tilde{\omega}_1^s,\tilde{\omega}_2^{\mathrm{T}})^{\mathrm{T}}],\\
R_n^*(\tilde{\omega}_1,\tilde{\omega}_2)&=n^{-1/(2s)}\tilde{\omega}_1\{T_n^*(\tilde{\omega}_1^s,\tilde{\omega}_2)+o_p(1)\},\\
T_n^*(\tilde{\omega}_1^s,\tilde{\omega}_2)&=(\tilde{\omega}_1^s,\tilde{\omega}_2^{\mathrm{T}})[\{n^{-1/2}L_n^{(s+1,0)}/(s+1)!,(n^{-1/2}L_n^{(1,1)})^{\mathrm{T}}\}^{\mathrm{T}}-C(\tilde{\omega}_1^s,\tilde{\omega}_2^{\mathrm{T}})^{\mathrm{T}}],
\end{aligned} \tag{34}$$

and

$$I=\begin{bmatrix}I_{11}&I_{12}\\I_{21}&I_{22}\end{bmatrix},\qquad C=\begin{bmatrix}C_{11}&C_{12}\\C_{21}&C_{22}\end{bmatrix}. \tag{35}$$

The remainder $R_n^*(\tilde{\omega}_1, \tilde{\omega}_2)$ will be used later to calculate the statistic that asymptotically determines the sign of $\hat{\theta}_1 - \theta_1^*$ when s is even.

Part (a) of Theorems 3 and 4 follows because the regularity conditions (A1), specifically the compactness of Θ, (A2) and (A3), and the continuity of $\log f(y; \theta)$ guarantee the existence, uniqueness with a probability tending to 1 and consistency of the MLE $\hat{\theta}$ of θ when $\theta = \theta^*$ (Newey and McFadden 1993, Theorem 2.5). Furthermore, because under (A1) Θ contains an open neighbourhood $\mathcal{N}$ of θ^* and because $\log f(y; \theta)$ is differentiable in $\mathcal{N}$, the MLE is with a probability tending to 1 a solution of the score equation

$$[L_n^{(1,0)}(\theta), (L_n^{(0,1)}(\theta))^{\mathrm{T}}] = (0, 0) \tag{36}$$

(Newey and McFadden, 1993, Section 3.7).

Define $(\omega_1, \omega_2^{\mathrm{T}}) \equiv [n^{1/(2s)}(\theta_1 - \theta_1^*), n^{1/2}(\theta_2 - \theta_2^*)^{\mathrm{T}}]$ and $(\hat{\omega}_1, \hat{\omega}_2^{\mathrm{T}}) \equiv [n^{1/(2s)}(\hat{\theta}_1 - \theta_1^*), n^{1/2}(\hat{\theta}_2 - \theta_2^*)^{\mathrm{T}}]$, where $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2^{\mathrm{T}})$ is the MLE of $\theta = (\theta_1, \theta_2)$. Forming the Taylor expansions of $n^{-1/(2s)} L_n^{(1,0)}(\hat{\theta})$ and $L_n^{(0,1)}(\hat{\theta})$ around θ^* and analysing the convergence of each term of these expansions similarly to what was done for the log-likelihood expansion gives that $\hat{\theta}$ must solve

$$0 = M_{1n}(\omega_1, \omega_2) + n^{-1/(2s)} \left\{ \begin{bmatrix} \omega_1^s & \omega_2 \\ 0 & \omega_1 \end{bmatrix} M_{2n}(\omega_1^s, \omega_2) + o_p(1) \begin{bmatrix} P_1(\omega_1, \omega_2) \\ P_2(\omega_1, \omega_2) \end{bmatrix} \right\}, \tag{37}$$

where

$$M_{1n}(\omega_1, \omega_2) = \begin{bmatrix} \omega_1^{s-1} & 0 \\ 0 & 1 \end{bmatrix} \left\{ \begin{bmatrix} n^{-1/2} L_n^{(s,0)}/s! \\ n^{-1/2} L_n^{(0,1)} \end{bmatrix} - \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} \begin{bmatrix} \omega_1^s \\ \omega_2 \end{bmatrix} \right\},$$

$$M_{2n}(\omega_1^s, \omega_2) = \begin{bmatrix} n^{-1/2} L_n^{(s+1,0)}/(s+1)! \\ n^{-1/2} L_n^{(1,1)} \end{bmatrix} - \begin{bmatrix} (2s+1)C_{11} & sC_{12} \\ C_{21} & C_{22} \end{bmatrix} \begin{bmatrix} \omega_1^s \\ \omega_2 \end{bmatrix},$$

and $P_1(\omega_1, \omega_2)$ and $P_2(\omega_1, \omega_2)$ are polynomials in ω_1 and ω_2. (Consistency of $\hat{\theta}$ under $\theta = \theta^*$ and assumption (A6) are used to show that the factor multiplying these polynomials is $o_p(1)$). By Corollary 1, $n^{-1/2} L_n^{(s,0)} = O_p(1)$, $n^{-1/2} L_n^{(0,1)} = O_p(1)$, $n^{-1/2} L_n^{(s+1,0)} = O_p(1)$, $n^{-1/2} L_n^{(1,1)} = O_p(1)$. Thus, since $\hat{\theta}$ satisfies (37), $\hat{\omega}_1 = O_p(1)$ and $\hat{\omega}_2 = O_p(1)$.

Define $(Z_{1n}, Z_{2n}^{\mathrm{T}})^{\mathrm{T}} \equiv I^{-1}[n^{-1/2} L_n^{(s,0)}/s!, (n^{-1/2} L_n^{(0,1)})^{\mathrm{T}}]^{\mathrm{T}}$, where the matrix I is defined as in Theorems 3 and 4. The matrix I^{-1} exists since, by assumption (B2), I is non-singular. By definition, $Z_{1n} = \Delta^{-1} \times (n^{-1/2} L_n^{(s,0)}/s! - I_{12} I_{22}^{-1} n^{-1/2} L_n^{(0,1)})$ and $Z_{2.1n} = I_{22}^{-1} n^{-1/2} L_n^{(0,1)}$, where $Z_{2.1n} \equiv Z_{2n} - I^{21}(I^{11})^{-1} Z_{1n}$, $\Delta \equiv (I_{11} - I_{12} I_{22}^{-1} I_{21})$ and I^{jk} is the (j, k)th entry of the matrix I^{-1}. By Corollary 1(a), $(Z_{1n}, Z_{2n}^{\mathrm{T}})^{\mathrm{T}} = (Z_1, Z_2^{\mathrm{T}})^{\mathrm{T}} + o_p(1)$, where $(Z_1, Z_2^{\mathrm{T}})^{\mathrm{T}}$ $N(0, I^{-1})$. Solving for $\hat{\omega}_2$ in the second equation of (37) and substituting its solution in the first equation of (37) and in (33) gives

$$0 = \hat{\omega}_1^{s-1}(Z_{1n} - \hat{\omega}_1^s) + O_p(n^{-1/(2s)}), \tag{38}$$

and

$$L_n(\hat{\theta}) = L_n(\theta^*) + \hat{\omega}_1^s \Delta (Z_{1n} - \hat{\omega}_1^s/2) + Z_{2.1n}^{\mathrm{T}} I_{22} Z_{2.1n} + O_p(n^{-1/(2s)}). \tag{39}$$

Next we show that when s is odd, (38) and (39) imply that

$$\hat{\omega}_1^s = Z_{1n} + O_p(n^{-1/(2s)}). \tag{40}$$

But when (40) holds, substituting $\hat{\omega}_1^s$ with $Z_{1n} + O_p(n^{-1/(2s)})$ in the second equation of (37) implies that

$$(\hat{\omega}_1^s,\ \hat{\omega}_2^{\mathrm{T}}) = (Z_{1n},\ Z_{2n}^{\mathrm{T}}) + O_p(n^{-1/(2s)}), \tag{41}$$

which shows Theorem 3(b). To show (40) it is enough, by (38), to show that there exists no subsequence n' of n such that $\hat{\omega}_1$ converges in law along the subsequence to a random variable with an atom of probability at 0. That is, if $\hat{\omega}_1 \rightsquigarrow U$ along the subsequence n', then for all $\delta > 0$, there exists an $\varepsilon > 0$ such that $P(|U| > \varepsilon) > 1 - \delta$. We show this by contradiction. Henceforth, suppose that there exists a subsequence n' such that $\hat{\omega}_1 \rightsquigarrow U$ and $P(U = 0) = \delta > 0$. Thus, by (39), there exists a constant M such that for all $\varepsilon > 0$, $L_{n'}(\hat{\theta}) - L_{n'}(\theta^*) < \varepsilon M + Z_{2.1n'}^{\mathrm{T}} I_{22} Z_{2.1n'}^{\mathrm{T}}/2$ with probability that converges to a number greater that $\delta/2$ along the subsequence. However, letting $\bar{\theta} = (\overline{\omega}_1, \overline{\omega}_2)$, where $\overline{\omega}_1^s = Z_{1n'}$ and $\overline{\omega}_2 = I_{22}^{-1}(n^{-1/2} L_{n'}^{(0,1)} - I_{21}\overline{\omega}_1^s)$, we have that

$$L_{n'}(\bar{\theta}) - L_{n'}(\theta^*) = \Delta Z_{1n'}^2/2 + Z_{2.1n'}^{\mathrm{T}} I_{22} Z_{2.1n'}/2 + O_p(n'^{-1/(2s)}).$$

Thus, $L_{n'}(\bar{\theta}) - L_{n'}(\hat{\theta}) > \Delta Z_{1n'}^2/4 > \sigma > 0$ with probability converging along the subsequence to a strictly positive number. This is a contradiction since $\hat{\theta}$ is the MLE. This concludes the proof of Theorem 3(b). Next, by (41), evaluating (33) at $(\hat{\omega}_1, \hat{\omega}_2)$ gives

$$L_n(\hat{\theta}) = L_n(\theta^*) + \tfrac{1}{2}(Z_1,\ Z_2^{\mathrm{T}}) I (Z_1,\ Z_2^{\mathrm{T}})^{\mathrm{T}} + O_p(n^{-1/(2s)}). \tag{42}$$

This shows Theorem 3(c) since $(Z_1,\ Z_2^{\mathrm{T}}) \sim N(0,\ I^{-1})$.

To show Theorem 4(b), note that since $\hat{\omega}_1^s > 0$ when s is even, then, conditional on $Z_{1n} < 0$, $Z_{1n} - \hat{\omega}_1^s < Z_{1n} < 0$ and therefore conditional on $Z_{1n} < 0$, $Z_{1n} - \hat{\omega}_1^s$ cannot converge to a random variable with an atom of probability at 0 along any subsequence. Thus, by (38), conditional on $Z_{1n} < 0$, $\hat{\omega}_1 = o_p(1)$. So, by (37), conditional on $Z_{1n} < 0$,

$$(\hat{\omega}_1,\ \hat{\omega}_2^{\mathrm{T}}) - [0,\ (I_{22}^{-1} n^{-1/2} L_n^{(0,1)})^{\mathrm{T}}] = O_p(n^{-1/(2s)}), \tag{43}$$

or equivalently $(\hat{\omega}_1,\ \hat{\omega}_2^{\mathrm{T}}) = (0,\ Z_{2.1n}^{\mathrm{T}}) + O_p(n^{-1/(2s)})$. For $Z_{1n} > 0$, arguing as for s odd, we conclude that $\hat{\omega}_1$ cannot converge to a random variable with an atom of probability at 0 along any subsequence, and hence $(\hat{\omega}_1, \hat{\omega}_2)$ must satisfy (41). Thus, conditional on $Z_{1n} > 0$, $(|\hat{\omega}_1|, \hat{\omega}_2^{\mathrm{T}}) = (Z_{1n}^{1/s},\ Z_{2n}^{\mathrm{T}}) + O_p(n^{-1/(2s)})$. To calculate the statistic that asymptotically determines the sign of $\hat{\omega}_1$, we note that by (33) the log-likelihood evaluated at $(\hat{\omega}_1, \hat{\omega}_2)$ depends on the sign of $\hat{\omega}_1$ only through the remainder $R_n^*(\hat{\omega}_1, \hat{\omega}_2)$. Thus, the sign of $\hat{\omega}_1$ must be chosen to maximize this remainder. But, by (34), $R_n^*(\hat{\omega}_1, \hat{\omega}_2) = n^{-1/(2s)}\hat{\omega}_1\{\hat{T}_n^* + o_p(1)\}$, where

$$\hat{T}_n^* = [n^{-1/2} L_n^{(s,0)}/(s)!,\ (n^{-1/2} L_n^{(0,1)})^{\mathrm{T}}] I^{-1} \times \left\{ \begin{bmatrix} n^{-1/2} L_n^{(s+1,0)}/(s+1)! \\ n^{-1/2} L_n^{(1,1)} \end{bmatrix} - CI^{-1} \begin{bmatrix} n^{-1/2} L_n^{(s,0)}/(s)! \\ n^{-1/2} L_n^{(0,1)} \end{bmatrix} \right\}. \tag{44}$$

Thus, $P(\hat{\omega}_1 \hat{T}_n^* > 0) \to 1$ as $n \to \infty$. Equivalently, $P\{I(\hat{\omega}_1 > 0) = I(\hat{T}_n^* > 0)\} \to 1$ as $n \to \infty$. But,

$$\begin{bmatrix} n^{-1/2} L_n^{(s+1,0)}/(s+1)! \\ n^{-1/2} L_n^{(1,1)} \end{bmatrix} - CI^{-1} \begin{bmatrix} n^{-1/2} L_n^{(s,0)}/(s)! \\ n^{-1/2} L_n^{(0,1)} \end{bmatrix} = n^{-1/2} \sum_{i=1}^{n} D_i,$$

where D_i is, by Corollary 1, the residual from the population regression of the vector $[l^{[s+1,0]}(Y_i; \theta^*), (l^{[1,1]}(Y_i; \theta^*))^{\mathrm{T}}]$ on the vector $[l^{[s,0]}(Y_i; \theta^*), (l^{[0,1]}(Y_i; \theta^*))^{\mathrm{T}}]$, which is not identically equal to zero by assumption (B3). Thus, D_i is uncorrelated with $[l^{[s,0]}(Y_i; \theta^*), (l^{[0,1]}(Y_i; \theta^*))^{\mathrm{T}}]$. Also,

$$\sum_i [l^{[s,0]}(Y_i; \theta^*), (l^{[0,1]}(Y_i; \theta^*))^{\mathrm{T}}] = [L_n^{(s,0)}/(s)!, (L_n^{(0,1)})^{\mathrm{T}}].$$

Thus, $n^{-1/2} \sum_{i=1}^{n} D_i \rightsquigarrow (W_1, W_2^{\mathrm{T}})$, with (W_1, W_2^{T}) a mean-zero normal random vector uncorrelated with (Z_1, Z_2^{T}). Finally, $\hat{T}_n^* \rightsquigarrow Z_1 W_1 + Z_2^{\mathrm{T}} W_2$ and $P(Z_1 W_1 + Z_2^{\mathrm{T}} W_2 > 0 | Z_1, Z_2) = \frac{1}{2}$. Thus, $B \equiv I(Z_1 W_1 + Z_2^{\mathrm{T}} W_2 > 0)$ is independent of (Z_1, Z_2^{T}). This concludes the proof of Theorem 4(b). Theorem 4(c) then follows by noting that, conditional on $Z_{1n} < 0$, (43) holds, and then, from (39),

$$L_n(\hat{\theta}) = L_n(\theta^*) + Z_{2.1n}^{\mathrm{T}} I_{22} Z_{2.1n}/2 + O_p(n^{-1/(2s)}),$$

and conditional on $Z_{1n} > 0$, (41) holds and therefore (42) holds.

The proof of Theorems 3 and 4 for an arbitrary p follows identically as for the case $p = 2$ after making the following series of substitutions: first,

$$\theta_2 \to (\theta_2, \theta_3, \ldots, \theta_p), \qquad \theta_2^* \to (\theta_2^*, \theta_3^*, \ldots, \theta_p^*), \qquad \tilde{\theta}_2 \to (\tilde{\theta}_2, \tilde{\theta}_3, \ldots, \tilde{\theta}_p),$$
$$\hat{\theta}_2 \to (\hat{\theta}_2, \hat{\theta}_3, \ldots, \hat{\theta}_p);$$

second, for $s \leqslant j \leqslant 2s+1$, and with $r_{(j)}$ and $L_n^{(r)}(\theta)$ defined as in Section 5,

$$L_n^{(j,0)} \to L_n^{(r_{(j)})}(\theta^*),$$

and

$$\delta_n^{(2s+1,0)} \to L_n^{(r_{(2s+1)})}(\bar{\theta}) - L_n^{(r_{(2s+1)})}(\theta^*);$$

third, for $0 \leqslant j \leqslant 2s$, $2 \leqslant k \leqslant p$ and with $\tilde{R}_{(j,k)}$ defined as the $1 \times p$ vector with first entry equal to j, kth entry equal to 1 and all other entries equal to 0,

$$L_n^{(j,1)} \to [L_n^{(\tilde{R}_{(j,2)})}(\theta^*), L_n^{(\tilde{R}_{(j,3)})}(\theta^*), \ldots, L_n^{(\tilde{R}_{(j,p)})}(\theta^*)]^{\mathrm{T}}$$

and

$$\delta_n^{(2s,1)} \to [L_n^{(\tilde{R}_{(2s,2)})}(\bar{\theta}), L_n^{(\tilde{R}_{(2s,3)})}(\bar{\theta}), \ldots, L_n^{(\tilde{R}_{(2s,p)})}(\bar{\theta})]^{\mathrm{T}} - [L_n^{(\tilde{R}_{(2s,2)})}(\theta^*), L_n^{(\tilde{R}_{(2s,3)})}(\theta^*), \ldots, L_n^{(\tilde{R}_{(2s,p)})}(\theta^*)]^{\mathrm{T}};$$

fourth, for $0 \leqslant j \leqslant 2s-1$,

$$L_n^{(j,2)} \to C_{(j,2)n}(\theta^*)$$

and

$$\delta_n^{(2s-1,2)} \to C_{(2s-1,2)n}(\bar{\theta}) - C_{(2s-1,2)n}(\theta^*),$$

where $C_{(j,2)n}(\theta)$ is the $(p-1)\times(p-1)$ matrix with entry (u, v) equal to $\partial L_n^{j+2}(\theta)/\partial\theta_1^j\partial\theta_{u+1}\partial\theta_{v+1}$; and finally

$$n^{-1/(2s)}\tilde{\omega}_2^{\mathrm{T}}A_{18n} \to n^{-1/(2s)}\sum_{k=3}^{2s+1}\sum_{\substack{r.=k\\ r.-r_1\geqslant 3}} c_r n^{-1}L_n^{(r)}(\theta^*)n^{\{(1-r_1)+(2-r.+r_1)s\}/(2s)}\omega_1^{r_1}\omega_2^{r_2}\ldots\omega_p^{r_p};$$

$$n^{-1/(2s)}\tilde{\omega}_2^{\mathrm{T}}A_{19n} \to n^{-1/(2s)}\sum_{\substack{r.=2s+1\\ r.-r_1\geqslant 3}} c_r n^{-1}\delta_n^{(r)}n^{\{(1-r_1)+(2-r.+r_1)s\}/(2s)}\omega_1^{r_1}\omega_2^{r_2}\ldots\omega_p^{r_p};$$

where c_r are appropriate constants, and, for any r such that $r. = 2s+1$ and $r.-r_1 \geqslant 3$, $\delta_n^{(r)}$ is defined as $L_n^{(r)}(\theta^*) - L_n^{(r)}(\bar{\theta})$.

Furthermore, I_{12}, I_{21} and I_{22} are redefined respectively as the $1\times(p-1)$, $(p-1)\times 1$ and $(p-1)\times(p-1)$ block submatrices of the partitioned matrix I in (35), and C_{11}, C_{12}, C_{21}, C_{22} are redefined respectively as the 1×1, $1\times(p-1)$, $(p-1)\times 1$ and $(p-1)\times(p-1)$ block submatrices of the partitioned matrix in (35), where I is the matrix defined in Theorems 3 and 4 and C is defined as

$$C \equiv \mathrm{E}\{[l^{(r_{(s+1)})}(Y;\theta^*)/(s+1)!,\ l^{(\tilde{R}_{(1,2)})}(Y;\theta^*),\ \ldots,\ l^{(\tilde{R}_{(1,p)})}(Y;\theta^*)]^{\mathrm{T}}$$

$$\times[l^{(r_{(s)})}(Y;\theta^*)/s!,\ l^{(\tilde{R}_{(0,2)})}(Y;\theta^*),\ \ldots,\ l^{(\tilde{R}_{(0,p)})}(Y;\theta^*)]\}.$$

Derivation of the results in Section 3.3

Derivations for the case $b \geqslant 1/(2s)$

In the proof of Theorems 1–4, we showed that under the regularity conditions (A1)–(A7) of Section 4.2, equation (4) holds for any θ such that $n^{1/2}(\theta-\theta^*)^s = O(1)$ when the data are generated under θ^*. This implies that $L_n(\theta_n) - L_n(\theta^*) = G_n(\theta_n) + o_p(1)$. Thus, as $n\to\infty$, $L_n(\theta_n) - L_n(\theta^*)$ converges under θ^* to 0 when $b > 1/(2s)$ and to $Z_0a^s - Ia^{2s}/2$ when $b = 1/(2s)$. Hence, by LeCam's first lemma (Hájek and Šidák 1967, p. 202) the sequences of distributions with densities $f^n(y;\theta_n)$ and $f^n(y;\theta^*)$ are contiguous. Similarly, $f^n(y;\theta'_n)$ and $f^n(y;\theta^*)$ are contiguous. Thus, for any sequence of random variables X_n, $n = 1, 2, \ldots$,

$$X_n = o_p(1) \Rightarrow X_n = o_{p_n}(1). \tag{45}$$

This implies that equation (4) also holds for values of θ satisfying $n^{1/2}(\theta-\theta^*)^s = O(1)$ when the data are generated under θ_n. In addition, by LeCam's third lemma (Hájek and Šidák, 1967, p. 208), $Z_0 \sim N(a^sI, I)$ when $b = 1/(2s)$ and $Z_0 \sim N(0, I)$ when $b > 1/(2s)$. Remark (R3.1) follows from (45). Remark (R3.2) follows from the fact that $P(Z_0 < 0) = 1/2$ if $b > 1/(2s)$ and $P(Z_0 < 0) = \Phi(-a^s\sqrt{I})$ if $b = 1/(2s)$. To show (R3.3), notice that, by (45),

$$L_n(\hat{\theta}) - L_n(\theta^*) = G_n(\hat{\theta}) + o_{p_n}(1), \tag{46}$$

$$n^{1/2}(\hat{\theta} - \theta^*)^s = I^{-1} Z_0 I(Z_0 > 0) + o_{p_n}(1) \qquad \text{if } s \text{ is even} \tag{47}$$

and

$$n^{1/2}(\hat{\theta} - \theta^*)^s = I^{-1} Z_0 + o_{p_n}(1) \qquad \text{if } s \text{ is odd.} \tag{48}$$

Remark (R3.3) then follows by substituting $n^{1/2}(\hat{\theta} - \theta^*)^s$ in the expression for $G_n(\hat{\theta})$ with the right-hand side of expressions (47) and (48). The distribution of $2\{L_n(\hat{\theta}) - L_n(\theta^*)\}$ under θ'_n follows by an identical argument because, by contiguity of $f^n(y; \theta'_n)$ and $f^n(y; \theta^*)$, (46), (47) and (48) are also valid when $o_{p_n}(1)$ is used to indicate convergence to 0 in probability under θ'_n. To show (R3.4), notice that $L_n(\theta'_n) - L_n(\theta_n) = o_{p_n}(1)$ since expression (4) is valid under θ_n and $G_n(\theta)$ is symmetric around θ^*. Remark (R3.4) then follows by LeCam's third lemma.

Derivations for the case $b < 1/(2s)$

In what follows we assume that, in addition to (A1)–(A7) and (B1)–(B3), the following regularity conditions hold:

(D1) With probability 1, $l(Y; \theta)$ has $2s + 2$ derivatives with respect to θ, for all $\theta \in \mathcal{N}$.
(D2) For $0 \leqslant j \leqslant 2s + 2$ and some $\nu > 2$, $\sup_{\theta \in \mathcal{N}} \mathrm{E}_\theta\{|l^{(j)}(Y; \theta^*)|^\nu\} < \infty$, where the subscript θ indicates expectation under the parameter θ.
(D3) For $0 \leqslant j \leqslant 2s + 2$, $\sup_{\theta \in \mathcal{N}} \mathrm{E}_\theta[\{l^{(j)}(Y; \theta)\}^2] < \infty$.
(D4) Condition (A6) of Section 4.2 holds with $r. = 2s + 2$, with $g(Y)$ satisfying $\sup_{\theta \in \mathcal{N}} \mathrm{E}_\theta\{g(Y)^2\} < \infty$.

Suppose first that $1/(2s + 2) \leqslant b < 1/(2s)$. We first show the identities (10) and (16). Let $\mu_{n,j} \equiv n^{1/2}\mathrm{E}_n\{l^{[j]}(Y; \theta^*)\}$, where the subscript n denotes expectation taken under $\theta = \theta_n$. Under (D2), by the central limit theorem, for $j = 0, 1$, $n^{-1/2} L_n^{[s+j]}(\theta^*) = Z_j + \mu_{n,j} + o_{p_n}(1)$, and for $0 \leqslant j \leqslant s - 1$, $n^{-1/2} L_n^{[s+j]}(\theta^*) = \mu_{n,j} + O_{p_n}(1)$, where $O_{p_n}(1)$ denotes a sequence that is bounded in probability under θ_n. Also under (D2), by the law of large numbers, $n^{-1} L_n^{[2s+j]}(\theta^*) = n^{-1/2}\mu_{n,2s+j} + o_{p_n}(1)$, $0 \leqslant j \leqslant 2$. Thus, with

$$d_{jn}(\theta) \equiv n^{1/2}\{(\theta - \theta^*)^{s+j} - (\theta_n - \theta^*)^{s+j}\}, \qquad j \geqslant 0,$$

a Taylor expansion of $L_n(\theta)$ and $L_n(\theta_n)$ around θ^* gives

$$
\begin{aligned}
L_n(\theta) = L_n(\theta_n) &- \left[\sum_{j=0}^{1}\{Z_j + \mu_{n,s+j} + o_{p_n}(1)\}d_{jn}(\theta)\right] \\
&+ \left[\sum_{j=0}^{1}\{n^{-1/2}\mu_{n,2s+j} + o_{p_n}(1)\}n^{1/2}d_{s+j,n}(\theta)\right] \\
&+ \left[\sum_{j=2}^{s-1}\{\mu_{n,s+j} + O_{p_n}(1)\}d_{jn}(\theta)\right] + [O_{p_n}(1)n^{1/2}d_{s+2,n}(\theta)] \\
&+ \{\Delta_{1n}n(\theta-\theta^*)^{2s+2} + \Delta_{2n}n(\theta_n-\theta^*)^{2s+2}\} \\
&= L_n(\theta_n) - (B_{1n} + B_{2n} + \ldots + B_{5n}), \qquad (49)
\end{aligned}
$$

where $\Delta_{1n} = n^{-1}\{L_n^{[2s+2]}(\tilde{\theta}') - L_n^{[2s+2]}(\theta^*)\}$, $\Delta_{2n} = n^{-1}\{L_n^{[2s+2]}(\tilde{\theta}'') - L_n^{[2s+2]}(\theta^*)\}$ for some $\tilde{\theta}'$ and $\tilde{\theta}''$ satisfying $\|\tilde{\theta}' - \theta^*\| \leqslant \|\theta - \theta^*\|$ and $\|\tilde{\theta}'' - \theta^*\| \leqslant \|\theta_n - \theta^*\|$, and the terms B_{jn} correspond to the square-bracketed terms of the expansion in the order they appear. But under (D1)–(D3),

$$
\begin{aligned}
\mu_{n,s} &= In^{1/2}(\theta_n-\theta^*)^s + Cn^{1/2}(\theta_n-\theta^*)^{s+1} + o(1), \\
\mu_{n,s+1} &= Cn^{1/2}(\theta_n-\theta^*)^s + Jn^{1/2}(\theta_n-\theta^*)^{s+1} + o(1), \\
\mu_{n,s+j} &= O\{n^{1/2}(\theta_n-\theta^*)^s\}, \qquad 2 \leqslant j \leqslant 2s-1, \\
n^{-1/2}\mu_{n,2s} &= -I/2 + o(1), \qquad n^{-1/2}\mu_{n,2s+1} = -C + o(1), \qquad n^{-1/2}\mu_{n,2s+2} = O(1).
\end{aligned}
$$

Suppose first that $n^{1/2}\{(\theta-\theta^*)^s - (\theta_n-\theta^*)^s\} = O(1)$. Then with

$$
\begin{aligned}
c_{1jn}(\theta) &\equiv n^{1/2}(\theta_n-\theta^*)^s d_{jn}(\theta), \qquad j \geqslant 2, \\
c_{2n}(\theta) &\equiv n^{1/2}d_{s+2,n}(\theta), \\
c_{3n}(\theta) &\equiv n(\theta_n-\theta^*)^{2s+2},
\end{aligned}
$$

the following identities hold:

$$
c_{1jn}(\theta),\ c_{2n}(\theta),\ c_{3n}(\theta) = o(1)\ \ d_{1n}(\theta) \text{ if } 1/(2s+2) < b < 1/(2s), \qquad (50)
$$

$$
c_{1jn}(\theta),\ c_{2n}(\theta) = o(1) \qquad \text{if } b = 1/(2s+2), \qquad (51)
$$

$$
c_{3n}(\theta) = a^{2s+2} \qquad \text{if } b = 1/(2s+2). \qquad (52)
$$

Equation (50) implies that $B_{3n} = o_{p_n}(1)d_{1n}(\theta)$ if $1/(2s+2) < b < 1/(2s)$, and (51) implies that $B_{3n} = o_{p_n}(1)$ if $b = 1/(2s+2)$. Equation (51) implies that $B_{4n} = o_{p_n}(1)d_{1n}(\theta)$ if $1/(2s+2) < b < 1/(2s)$, and also implies that $B_{4n} = o_{p_n}(1)$ if $b = 1/(2s+2)$. Furthermore, under assumption (C4), $\Delta_{1n} = o_{p_n}(1)$ and $\Delta_{2n} = o_{p_n}(1)$, and therefore (51) implies that $B_{5n} = o_{p_n}(1)d_{1n}(\theta)$ if $1/(2s+2) < b < 1/(2s)$, and (51) and (52) imply that $B_{5n} = o_{p_n}(1)$ if

$b = 1/(2s+2)$. In addition, by $n^{1/2}(\theta_n - \theta^*)^{s+2} = o(1)$ and after some algebra, it can be seen that

$$\begin{aligned}
&\sum_{j=0}^{1} \mu_{n,s+j} d_{jn}(\theta) + n^{-1/2}\mu_{n,2s+j} n^{1/2} d_{s+j,n}(\theta) \\
&\quad = -\frac{I}{2} n\{(\theta-\theta^*)^s - (\theta_n-\theta^*)^s\}^2 + Cn^{1/2}\prod_{j=0}^{1}\{(\theta-\theta^*)^{s+j} - (\theta_n-\theta^*)^{s+j}\} \\
&\qquad + Jn^{1/2}(\theta_n-\theta^*)^{s+1} n^{1/2}\{(\theta-\theta^*)^{s+1} - (\theta_n-\theta^*)^{s+1}\} + o(1). \qquad (53)
\end{aligned}$$

Equation (16) follows after substituting in the expansion (49) the left-hand side of (53) with its right-hand side and noting that, for $b = 1/(2s+2)$, $n^{1/2}(\theta_n - \theta^*)^{s+1} = a^{s+1}$ and $n^{1/2}\{(\theta-\theta^*)^{s+1} - (\theta_n-\theta^*)^{s+1}\} = -d_s(\theta)2a^{s+1} + o(1)$. Equation (10) follows by performing the same substitution in the expansion (49) and noting that when $b > 1/(2s+2)$, $n^{1/2}(\theta_n - \theta^*)^{s+1} = o(1)$.

We now show that the MLE $\hat{\theta}$ satisfies equations (13) when s is odd and $1/(2s+2) \leqslant b < 1/(2s)$, and when s is even it is equal to one of the points $\tilde{\theta}_1$ or $\tilde{\theta}_2$ satisfying equation (14) when $1/(2s+2) < b < 1/(2s)$ and satisfying equations (18) or (19) when $b = 1/(2s+2)$. Under the regularity conditions (A1)–(A7) the MLE $\hat{\theta}$ is a consistent estimator of θ when $\theta = \theta_n$ and solves the score equation $L_n^{(1)}(\theta) = 0$. Forming the Taylor expansion of $L_n^{(1)}(\hat{\theta})$ around θ_0, we obtain

$$\begin{aligned}
L_n^{(1)}(\hat{\theta}) = 0 = \sqrt{n}(\hat{\theta}-\theta^*)^{s-1} \Bigg[&\left\{ \sum_{j=0}^{1} (s+j)\{Z_j + \mu_{n,s+j} + o_{p_n}(1)\}(\hat{\theta}-\theta^*)^j \right\} \\
&+ \left\{ \sum_{j=2}^{s-1} (s+j)\{\mu_{n,s+j} + O_{p_n}(1)\}(\hat{\theta}-\theta^*)^j \right\}. \\
&+ \left\{ \sum_{j=0}^{1} (2s+j)\{n^{-1/2}\mu_{n,2s+j} + o_{p_n}(1)\}\sqrt{n}(\hat{\theta}-\theta^*)^{s+j} \right\} \\
&+ \left\{ (2s+2)\frac{L_n^{[2s+j]}(\bar{\theta})}{n}\sqrt{n}(\hat{\theta}-\theta^*)^{s+2} \right\} \Bigg] \\
= \sqrt{n}(\hat{\theta}-\theta^*)^{s-1}&(B'_{1n} + B'_{2n} + B'_{3n} + B'_{4n}). \qquad (54)
\end{aligned}$$

The terms B'_{jn} correspond to terms in braces in the order they appear. Note, first, that $n^{1/\{2(s+1)\}}(\hat{\theta}-\theta^*)$ is equal to

$$\operatorname{sgn}(\theta_n-\theta^*)^s \operatorname{sgn}(\hat{\theta}-\theta^*)^{s+1} n^{1/\{2(s+1)\}} |\theta_n - \theta^*| [1 + n^{1/2}\{(\hat{\theta}-\theta^*)^s - (\theta_n-\theta^*)^s\} o(1)].$$

Thus, for $b > 1/(2s+2)$, $n^{s/\{2(s+1)\}}(\theta_n - \theta^*)^s = o(1)$, and we have

$$n^{1/\{2(s+1)\}}(\hat{\theta}-\theta^*) = o(1)[1 + n^{1/2}\{(\hat{\theta}-\theta^*)^s - (\theta_n-\theta^*)^s\}], \tag{55}$$

and for $b = 1/(2s+2)$, $n^{s/\{2(s+1)\}}(\theta_n-\theta^*)^s = a^s$ and we have

$$n^{1/\{2(s+1)\}}(\hat{\theta}-\theta^*) = \operatorname{sgn}(a^s)|a|\operatorname{sgn}(\hat{\theta}-\theta^*)^{s+1}[1 + n^{1/2}\{(\hat{\theta}-\theta^*)^s - (\theta_n-\theta^*)^s\}o(1)]. \tag{56}$$

Consider now the term B'_{2n}. For $j \geqslant 2$,

$$\begin{aligned}\mu_{n,s+j}(\hat{\theta}-\theta^*)^j &= \{n^{s/\{2(s+1)\}}(\theta_n-\theta^*)^s n^{1/\{2(s+1)\}}(\hat{\theta}-\theta^*)\}(\hat{\theta}-\theta^*)^{j-1} \\ &= [1 + n^{1/2}\{(\hat{\theta}-\theta^*)^s - (\theta_n-\theta^*)^s\}]o_{p_n}(1).\end{aligned} \tag{57}$$

where the last equality follows from equations (55) and (56) and the consistency of $\hat{\theta}$. Thus, B'_{2n} is also equal to the last member of (57). Turn now to the term B'_{4n}. By equations (55) and (56) and by $\sqrt{n}(\hat{\theta}-\theta^*)^{s+2} = \{n^{1/(2(s+1))}(\hat{\theta}-\theta^*)\}^{s+1}(\hat{\theta}-\theta^*)$ and the consistency of $\hat{\theta}$, we have that $\sqrt{n}(\hat{\theta}-\theta^*)^{s+2}$ is equal to the last member of equation (57). Also, by (D1)–(D4) and the consistency of $\hat{\theta}$, $n^{-1}L_n^{[2s+2]}(\bar{\theta}) = o_{p_n}(1)$. Thus, B'_{4n} is also equal to the last member of equation (57). Now, by the consistency of $\hat{\theta}$ and after some algebra,

$$\begin{aligned}B'_{1n} + B'_{3n} = s[&\{Z_0 + o_{p_n}(1)\} - \{I + o_{p_n}(1)\}n^{1/2}\{(\hat{\theta}-\theta^*)^s - (\theta_n-\theta^*)^s\} \\ &+ \{C + o_{p_n}(1)\}n^{1/2}\{(\theta_n-\theta^*)^{s+1} - (\hat{\theta}-\theta^*)^{s+1}\} + o_{p_n}(1)].\end{aligned}$$

For $1/(2s+2) < b < 1/(2s)$, we have by (55) that $n^{1/2}\{(\theta_n-\theta^*)^{s+1} - (\hat{\theta}-\theta^*)^{s+1}\} = [n^{1/2}\{(\hat{\theta}-\theta^*)^s - (\theta_n-\theta^*)^s\} + 1]o_{p_n}(1)$. Thus,

$$B'_{1n} + B'_{3n} = s[\{Z_0 + o_{p_n}(1)\} - \{I + o_{p_n}(1)\}n^{1/2}\{(\hat{\theta}-\theta^*)^s - (\theta_n-\theta^*)^s\} + o_{p_n}(1)],$$

and from (54) we conclude that $\hat{\theta}$ satisfies

$$\hat{\theta} = \theta^* \text{ or } n^{1/2}\{(\hat{\theta}-\theta^*)^s - (\theta_n-\theta^*)^s\} = I^{-1}Z_0 + o_{p_n}(1). \tag{58}$$

For $b = 1/(2s+2)$, we have by (56) that

$$\begin{aligned}B'_{1n} + B'_{3n} = s[&\{Z_0 + o_{p_n}(1)\} - \{I + o_{p_n}(1)\}n^{1/2}\{(\hat{\theta}-\theta^*)^s - (\theta_n-\theta^*)^s\} \\ &+ Ca^{s+1}\{1 - \operatorname{sgn}(\hat{\theta}-\theta^*)^{s+1}\} + o_{p_n}(1)],\end{aligned}$$

and from (54) we conclude that when s is odd or when s is even and $\operatorname{sgn}(\hat{\theta}-\theta^*) = 1$, $\hat{\theta}$ satisfies one of the identities in (58). When s is even, $\hat{\theta}$ satisfies either (58) or

$$n^{1/2}\{(\hat{\theta}-\theta^*)^s - (\theta_n-\theta^*)^s\} = I^{-1}\{Z_0 + 2a^{s+1}\} + o_{p_n}(1).$$

Thus, to show the desired identities it only remains to prove that the probability that $\hat{\theta}$ is equal to θ^* converges under $\theta = \theta_n$ to 0 as $n \to \infty$. But to show this it is enough to show that (7) holds. But (7) follows for $b < 1/(2s)$ by noticing that $n^{1/2}(\theta_n-\theta^*)^s$ diverges and under (D1)–(D4), a Taylor expansion of $L_n(\theta_n)$ around θ^* gives

$$L_n(\theta_n) = L_n(\theta^*) + n^{1/2}(\theta_n-\theta^*)^s\{Z_0 + (I + o_{p_n}(1))n^{1/2}(\theta_n-\theta^*)^s/2 + o_{p_n}(1)\},$$

which diverges to $+\infty$ because I is positive.

Remark (R5.8) made for the case $b = 1/(2s+2)$ follows by LeCam's third lemma

because, by equation (16), $2\{L_n(\theta'_n) - L_n(\theta_n)\}$ converges under θ_n to $-2a^{s+1}(Z_1 + a^{s+1}J)$, which is a normal random variable with mean equal to $-1/2$ times its variance.

Finally, we show that when s is odd and $b < 1/(2s)$ or when s is even and $b < 1/(2s+2)$, (8) holds. Because, by definition, $\theta'_n - \theta^* = -(\theta_n - \theta^*)$, a Taylor expansion of $L_n(\theta'_n)$ and $L_n(\theta_n)$ around θ^* gives

$$\begin{aligned} L_n(\theta'_n) &= L_n(\theta_n) - \left[\sum_{j=0}^{s-1}\{1-(-1)^{s+j}\}n^{-1/2}L_n^{[s+j]}(\theta^*)n^{1/2}(\theta'_n-\theta^*)^{s+j}\right] \\ &\quad + [2n^{-1}L_n^{[2s+1]}(\theta^*)n(\theta'_n-\theta^*)^{2s+1}] + [(\Delta_{1n}-\Delta_{2n})n(\theta'_n-\theta^*)^{2s+2}] \\ &= L_n(\theta_n) - (B''_{1n} + B''_{2n} + B''_{3n}), \end{aligned}$$

where $\Delta_{kn} = n^{-1}\{L_n^{(2s+2)}(\bar{\theta}_k) - L_n^{(2s+2)}(\theta^*)\}$ for some $\|\bar{\theta}_k - \theta^*\| \leqslant \|\theta_n - \theta^*\|$, $k = 1, 2$, and B''_{jn}, $j = 1, 2, 3$, correspond to the terms in square brackets in the order they appear. Under assumptions (D1)–(D4), $B''_{3n} = o_{p'_n}(1)n(\theta'_n - \theta^*)^{2s+1}$. When s is odd and $b < 1/(2s)$,

$$B''_{1n} = 2n^{1/2}(\theta'_n - \theta^*)^s[Z_0 + \{I + o_{p'_n}(1)\}n^{1/2}(\theta'_n - \theta^*)^s + o_{p'_n}(1)],$$

and $B''_{2n} = o_{p'_n}(1)n(\theta'_n - \theta^*)^{2s}$. Thus,

$$L_n(\theta'_n) - L_n(\theta_n) = \{I + o_{p'_n}(1)\}\{n^{1/2}(\theta'_n - \theta^*)^s\}^2 - \{Z_0 + o_{p'_n}(1)\}2n^{1/2}(\theta'_n - \theta^*)^s. \quad (59)$$

Then (59) converges in probability under θ'_n to $+\infty$ as $n \to \infty$ since I is positive and $n^{1/2}(\theta_n - \theta^*)^s$ diverges. When s is even and $b < 1/(2s+2)$, then

$$B''_{1n} = n^{1/2}(\theta'_n - \theta^*)^{s+1}[\{Z_1 + o_{p'_n}(1)\} + Cn^{1/2}(\theta'_n - \theta^*)^s + \{J + o_{p'_n}(1)\}n^{1/2}(\theta'_n - \theta^*)^{s+1}]$$

and

$$B''_{2n} = n(\theta'_n - \theta^*)^{2s+1}\{-C + o_{p'_n}(1)\}.$$

Thus,

$$L_n(\theta'_n) - L_n(\theta_n) = \{n^{1/2}(\theta'_n - \theta^*)^{s+1}\}^2\{J + o_{p'_n}(1)\} + n^{1/2}(\theta'_n - \theta^*)^{s+1}\{Z_1 + o_{p'_n}(1)\}.$$

And the convergence in probability of $L_n(\theta'_n) - L_n(\theta_n)$ to $+\infty$ under θ'_n follows since by definition J is positive and $n^{1/2}(\theta'_n - \theta^*)^{s+1}$ diverges when $b < 1/(2s+2)$.

Proof of Theorem 5

The proof of Theorem 5 will use the result of the following proposition that is easily shown by induction:

Proposition A.1. *Let $b(\psi): \mathbb{R} \to \mathbb{R}^{p\times 1}$ be a vector function of a scalar ψ whose jth derivative exists. Let $b^{(l)}(\psi)$* denote *$\partial^l b(\psi)/\partial\psi^l$, $1 \leqslant l \leqslant j$. Let $h(u): \mathbb{R}^{p\times 1} \to \mathbb{R}$ and assume that $h(u)$ is j times differentiable. Let $Dh(u)$ denote the gradient of $h(u)$. Define*

$m(\psi) = h \circ b(\psi)$ and let $m^{(j)}(\psi)$ denote the jth derivative of $m(\psi)$. Then there exists a function $H : \mathbb{R}^{1\times p(j-1)} \to \mathbb{R}$ such that

$$m^{(j)}(\psi) = b^{(j)}(\psi)^{\mathrm{T}} Dh\{b(\psi)\} + H\{b(\psi)^{\mathrm{T}}, b^{(1)}(\psi)^{\mathrm{T}}, \ldots, b^{(j-1)}(\psi)^{\mathrm{T}}\}.$$

Let $\psi(\theta) : \Theta \to \mathbb{R}^p$ be defined as $\psi(\theta) = \theta + \sum_{l=0}^{s-1} A_l(\theta_1 - \theta_1^*)^l$, and let Ψ denote the range of ψ. Define $\psi^* = \psi(\theta^*)$. Clearly, $\psi^* = \theta^*$. Furthermore, the first element of the vector $\psi(\theta)$ is equal to the first element of θ, i.e. $\psi_1(\theta) = \theta_1$. In addition, the function $\psi(\theta)$ is one-to-one and onto Ψ with inverse given by $\theta(\psi) = \psi - \sum_{l=0}^{s-1} A_l(\psi_1 - \psi_1^*)^l$. For any ψ in Ψ, let $\tilde{f}(Y; \psi)$ denote $f\{Y; \theta(\psi)\}$. Then clearly, for $2 \leqslant k \leqslant p$,

$$\left.\frac{\partial \log \tilde{f}(Y; \psi)}{\partial \psi_k}\right|_{\psi^*} = \left.\frac{\partial \log f(Y; \theta)}{\partial \theta_k}\right|_{\theta^*}, \tag{60}$$

and, by definition of $\tilde{S}_1^{(s+j)}$, $j = 0, 1$,

$$\left.\frac{\partial^{s+j} \log \tilde{f}(Y; \psi)}{\partial^{s+j} \psi_1}\right|_{\psi^*} = \tilde{S}_1^{(s+j)}. \tag{61}$$

Thus $\tilde{f}(Y; \psi)$ satisfies conditions (B2) and (B3) of Section 4.2 at the parameter ψ^*. We will later show that it also satisfies (B1). Then, since clearly $\tilde{f}(Y; \psi)$ satisfies all the other regularity conditions of Theorem 2, the asymptotic distribution of the MLE $\hat{\psi}$ of ψ^* and of the likelihood ratio test statistic follow from this theorem. But since (60) and (61) hold, the conclusions of Theorem 5 follow because: (a) by the invariance of the MLE, $\hat{\psi} = \psi(\hat{\theta})$; and (b) by $\psi^* = \psi(\theta^*)$ and the fact that $\tilde{\mathscr{P}} = \{\tilde{f}(Y; \psi) : \psi \in \Psi\}$ and $\mathscr{P} = \{f(Y; \theta) : \theta \in \Theta\}$ are the same statistical model, the likelihood ratio test statistics for testing $H_0 : \psi = \psi^*$ versus $H_1 : \psi \neq \psi^*$ in $\tilde{\mathscr{P}}$ and for testing $H_0 : \theta = \theta^*$ versus $H_1 : \theta \neq \theta^*$ in $\mathscr{P}$ are exactly the same. It remains to show that $\tilde{f}(Y; \psi)$ satisfies condition (B1). We show this by induction in s. For $s = 2$, (B1) is true because

$$\left.\frac{\partial \log \tilde{f}(Y; \psi)}{\partial \psi_1}\right|_{\psi=\psi^*} = \left.\frac{\partial \log f(Y; \psi - A_1(\psi_1 - \psi_1^*))}{\partial \psi_1}\right|_{\psi=\psi^*} = S_1 - K_1^{\mathrm{T}}\Gamma = 0,$$

where the second identity is true by (31). Suppose now that $\tilde{f}(Y; \psi)$ satisfies (B1) for $s - 1$. In order to show that $\tilde{f}(Y; \psi)$ also satisfies (B1) for s, it will be convenient to define

$$b(\psi_1) \equiv [\psi_1, \psi_2^*, \ldots, \psi_p^*]^{\mathrm{T}} - \sum_{l=0}^{s-1} A_l(\psi_1 - \psi_1^*)^l,$$

$$c(\psi_1) \equiv b(\psi_1) + A_{s-1}(\psi_1 - \psi_1^*)^{s-1},$$

$$h(u) \equiv \log f(Y; u),\ m(\psi_1) \equiv h\{b(\psi_1)\}.$$

With these definitions and letting the superscript (j) denote the jth derivative of a function, the following identities hold:

$$b(\psi_1^*) = c(\psi_1^*) = \psi^*, \tag{62}$$

$$b^{(j)}(\psi_1^*) = c^{(j)}(\psi_1^*), \qquad j = 1, \ldots, s-2, \tag{63}$$

$$c^{(s-1)}(\psi_1) = 0, \tag{64}$$

$$b^{(s-1)}(\psi_1) = -[0,\, K_{s-1}]^{\mathrm{T}}, \qquad m(\psi_1) = \log \tilde{f}\{Y;\, \psi_1,\, \psi_2^*, \ldots, \psi_p^*\}, \tag{65}$$

and, for any j,

$$m^{(j)}(\psi_1)|_{\psi_1=\psi_1^*} = \left.\frac{\partial^j \log \tilde{f}(Y;\, \psi)}{\partial \psi_1^j}\right|_{\psi=\psi^*}. \tag{66}$$

Now, letting $Dh(u)$ denote the gradient of $h(u)$, by Proposition A.1 there exist functions H_j, for $j = 1, \ldots, s-2$, such that

$$m^{(j)}(\psi_1)|_{\psi_1=\psi_1^*} = b^{(j)}(\psi_1^*)^{\mathrm{T}} Dh\{b(\psi_1^*)\} + H_j\{b(\psi_1^*),\, b^{(1)}(\psi_1^*), \ldots, b^{(j-1)}(\psi_1^*)\}; \tag{67}$$

but, by (62) and (63), the right-hand side of (67) is equal to

$$c^{(j)}(\psi_1^*)^{\mathrm{T}} Dh\{c(\psi_1^*)\} + H_j\{c(\psi_1^*),\, c^{(1)}(\psi_1^*), \ldots, c^{(j-1)}(\psi_1^*)\} = \mathrm{d}^j h\{c(\psi_1)\}/\mathrm{d}\psi_1^j|_{\psi_1=\psi_1^*}.$$

This in turn is equal to 0 by the inductive hypothesis. Thus, by (66) the first $s-2$ partial derivatives of $\log \tilde{f}(Y;\, \psi)$ with respect to ψ_1 evaluated at ψ^* are equal to 0. Also,

$$m^{(s-1)}(\psi_1)|_{\psi_1=\psi_1^*} = -[0,\, K_s] Dh\{c(\psi_1^*)\} + H_{s-1}\{c(\psi_1^*),\, c^{(1)}(\psi_1^*), \ldots, c^{(s-2)}(\psi_1^*)\}$$

$$= -K_{s-1}(S_2,\, S_3, \ldots, S_p)^{\mathrm{T}} + \left.\frac{\mathrm{d}^{s-1} h\{c(\psi_1)\}}{d\psi_1^{s-1}}\right|_{\psi_1=\psi_1^*} = 0,$$

where the first equality is by Proposition A.1, and equations (62), (63) and (65), the second equality is by Proposition A.1 and (64) and the third equality follows because, by assumption, (31) is true for $j = s-1$. This concludes the proof of the theorem.

Proof of Lemma 1

The lemma follows directly from Faà di Bruno's formula on the qth partial derivative of the composition of two functions. This formula, applied to $G(u) = \log h(u)$, gives

$$\frac{\partial^q G(u)}{\partial u_1^q} = \sum \frac{q!}{k_1! \ldots k_q!} \left(\frac{\mathrm{d}^p \log y}{\mathrm{d} y^p}\right) \left(\frac{\partial h(u)}{\partial u_1}\right)^{k_1} \left(\frac{1}{2!}\frac{\partial^2 h(u)}{\partial u_1^2}\right)^{k_2} \cdots \left(\frac{1}{q!}\frac{\partial^q h(u)}{\partial u_1^q}\right)^{k_q}, \tag{68}$$

where the summation extends over all partitions of q such that $p = \sum_{l=1}^q k_l$ and $q = \sum_{l=1}^q l k_l$.

Lemma 1(i) follows immediately from formula (68). Specifically, if $h^*_{r_{(j)}} = 0$ for all $1 \leqslant j \leqslant s-1$, then, for $1 \leqslant q \leqslant s-1$, all terms in the summation are equal to 0 and therefore $G^{(r_{(q)})}(u^*) = 0$. The proof that $G^{(r_{(j)})}(u^*) = 0$ for all $1 \leqslant j \leqslant s-1$ implies that $h^*_{r_{(j)}} = 0$ for all $1 \leqslant j \leqslant s-1$ is easily carried out by induction, the induction step consisting of noting that, under the inductive hypothesis, formula (68) implies that $G^{(r_{(q)})}(u^*) - G^{(r_{(q-1)})}(u^*) = h^*_{r_{(q)}}$. Part (ii) follows by noting that when $h^*_{r_{(j)}} = 0$, for all

$1 \leqslant j \leqslant s-1$, the only non-zero term in the summation in (68) when $s \leqslant q \leqslant 2s-1$ corresponds to the choice $k_j = 0$, $1 \leqslant j < q$ and $k_q = 1$. Part (iii) follows by noting that when $q = 2s$ there are only two non-zero terms in the summation in formula (68) which correspond to the choices (a) $k_j = 0$, $1 \leqslant j < 2s-1$, and $k_{2s} = 1$ and (b) $k_l = 0$, for $1 \leqslant l \leqslant 2s$, $l \neq s$ and $k_s = 2$. Part (iv) follows similarly by noting that when $q = 2s+1$, the only two non-zero terms in the summation in (68) correspond to the choices (a) $k_j = 0$, $1 \leqslant j < 2s$, and $k_{2s+1} = 1$ and (b) $k_l = 0$ for $1 \leqslant l \leqslant 2s$, $l \neq s, s+1$, and $k_s = k_{s+1} = 1$. Parts (v)–(ix) can be shown similarly by taking one or two derivatives with respect to u_2 in both sides of (68) and examining the non-zero terms in the summation.

Acknowledgements

Andrea Rotnitzky, Matteo Bottai and James Robins were partially supported by grants nos 2 R01 GM 48704-06 and 2 R01 AI 32475-07 of the US National Institutes of Health.

References

Barndorff-Nielsen, O.E. and Cox, D.R. (1989) *Asymptotic Techniques for Use in Statistics*. London: Chapman & Hall.

Bhattacharyya, A. (1946) On some analogues to the amount of information and their uses in statistical estimation. *Sankhyā*, **8**, 1–14, 201–208, 277–280.

Bickel, P., Klaassen, C.A.J., Ritov, J. and Wellner, J.A. (1993) *Efficient and Adaptive Inference in Semiparametric Models*. Baltimore, MD: Johns Hopkins University Press.

Chant, D. (1974) On asymptotic tests of composite hypotheses in non-standard conditions. *Biometrika*, **61**, 291–298.

Chen, J. (1995) Optimal rate of convergence for finite mixture models. *Ann. Statist.*, **23**, 221–233.

Chernoff, H. (1954) On the distribution of the likelihood ratio. *Ann. Math. Statist.*, **25**, 573–578.

Copas, J.B. and Li, H.G. (1997) Inference for non-random samples (with discussion). *J. Roy. Statist. Soc. Ser. B*, **59**, 55–95.

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

Cox, D.R. and Reid, N. (1987a) Approximations to noncentral distributions. *Canad. J. Statist.*, **15**, 105–114.

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

Diggle, P. and Kenward, M.G. (1994) Informative drop-out in longitudinal data analysis (with discussion). *Appl. Statist.*, **43**, 49–93.

Faà di Bruno, F. (1859) *Théorie Générale de l'Élimination*. Paris: De Leiber et Faraquet.

Geyer, C. J. (1994) On the asymptotics of constrained M-estimation. *Ann. Statist.*, **22**, 1993–2010.

Gronau, R. (1974) Wage comparisons: a selectivity bias. *J. Political Economy*, **82**, 1119–1143.

Hájek, J. and Šidák, Z. (1967) *Theory of Rank Tests*. New York: Academic Press.

Heckman, J.J. (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Ann. Econom. Social Measurement*, **5**, 475–492.

Ibragimov, I.A. and Has'minskii, R.Z. (1981) *Statistical Estimation, Asymptotic Theory*. New York: Springer-Verlag.

Kiefer, N.M. (1982) A remark on the parameterization of a model for heterogeneity. Working paper no 278. Department of Economics, Cornell University, Ithaca, NY.

Lee, L.F. (1981) A specification test for normality in the generalized censored regression models. Discussion paper no 81-146. Center for Economic Research, University of Minnesota, Minneapolis.

Lee, L.F. (1993) Asymptotic distribution of the maximum likelihood estimator for a stochastic frontier function model with a singular information matrix. *Econometric Theory*, **9**, 413–430.

Lee, L.F. and Chesher, A. (1986) Specification testing when score test statistic are identically zero. *J. Econometrics*, **31**, 121–149.

Moran, P.A.P. (1971) Maximum likelihood estimators in non-standard conditions. *Proc. Cambridge Philos. Soc.*, **70**, 441–450.

Nelson, F.D. (1977) Censored regression models with unobserved stochastic censoring thresholds, *J. Econometrics*, **6**, 309–327.

Newey, W. and McFadden, D. (1993) Estimation in large samples. In D. McFadden and R. Engler (eds), *Handbook of Econometrics*, Vol. 4. Amsterdam: North-Holland.

Rubin, D. B. (1976) Inference with missing data. *Biometrika*, **63**, 581–592.

Rothenberg, T.J. (1971) Identification in parametric models. *Econometrica*, **39**, 577–592.

Sargan, J.D. (1983) Identification and lack of identification. *Econometrica*, **51**, 1605–1633.

Self, S. and Liang, K.Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *J. Amer. Statist. Assoc.*, **82**, 605–610.

Silvey, S.D. (1959) The Lagrangean multiplier test. *Ann. Math. Statist.*, **30**, 389–407.

Received August 1997 and revised November 1998