A DECOMPOSITION FOR THE LIKELIHOOD RATIO STATISTIC AND THE BARTLETT CORRECTION—A BAYESIAN ARGUMENT

BY PETER J. BICKEL[1,2] AND J. K. GHOSH[2]

University of California, Berkeley and Indian Statistical Institute

Let $l(\theta) = n^{-1} \log p(x, \theta)$ be the log likelihood of an n-dimensional X under a p-dimensional θ. Let $\hat{\theta}_j$ be the mle under H_j: $\theta^1 = \theta_0^1, \ldots, \theta^j = \theta_0^j$ and $\hat{\theta}_0$ be the unrestricted mle. Define T_j as

$$\left[2n\left\{l\left(\hat{\theta}_{j-1}\right) - l\left(\hat{\theta}_j\right)\right\}\right]^{1/2} \operatorname{sgn}\left(\hat{\theta}_{j-1}^j - \theta_0^j\right).$$

Let $T = (T_1, \ldots, T_p)$. Then under regularity conditions, the following theorem is proved: Under $\theta = \theta_0$, T is asymptotically $N(n^{-1/2}a_0 + n^{-1}a, J + n^{-1}\Sigma) + O(n^{-3/2})$ where J is the identity matrix. The result is proved by first establishing an analogous result when θ is random and then making the prior converge to a degenerate distribution. The existence of the Bartlett correction to order $n^{-3/2}$ follows from the theorem. We show that an Edgeworth expansion with error $O(n^{-2})$ for T involves only polynomials of degree less than or equal to 3 and hence verify rigorously Lawley's (1956) result giving the order of the error in the Bartlett correction as $O(n^{-2})$.

1. Introduction. Let $X = (X_1, \ldots, X_n)$ be a vector of observations with joint density $p(x, \theta)$, $\theta \in \Theta$ open $\subset R^p$, where we do not assume a priori any particular structure on $p(x, \theta)$. Consider the hypothesis H: $\theta^1 = \theta_0^1, \ldots, \theta^k = \theta_0^k$. Suppose that maximum likelihood estimates $\hat{\theta}$ and $\hat{\theta}_H$ for $\theta \in \Theta$ and $\theta \in H$, respectively, are well defined. Then let

$$l(\theta) = n^{-1} \log p(X, \theta), \tag{1.1}$$

$$l(\hat{\theta}) = \max_{\Theta} l(\theta), \tag{1.2}$$

$$l(\hat{\theta}_H) = \max_H l(\theta) \tag{1.3}$$

and

$$\Lambda = 2n\left(l(\hat{\theta}) - l(\hat{\theta}_H)\right) \tag{1.4}$$

the usual likelihood ratio test statistic. All these quantities, of course, depend on n but we suppress this dependence to ease the notation. There is a common approximation to the distribution of Λ which has the status of a folk theorem:

$$L_\theta(\Lambda) \approx \chi_k^2$$

Received September 1987; revised July 1989.

[1]This paper was completed while the author was visiting AT & T Bell Telephone Labs, the Courant Institute and the University of Chicago.

[2]Research partially supported by ONR Contract N00014-80-C-0163.

AMS 1980 *subject classifications.* 62F05, 62F15.

Key words and phrases. Bartlett correction, signed log likelihood ratio statistic, Bernstein–von Mises theorem.

for $\theta \in H$. Theoretically this can be interpreted, for $\theta \in H$, as

$$P_\theta[\Lambda \le t] = \chi_k^2(t) + o(1) \tag{1.5}$$

as $n \to \infty$. This result was proved by Wilks (1938) and extended by Wald (1943) in the i.i.d. case, extended to the Markov case by Billingsley (1961) and subsequently extended to many other dependent and nonstationary situations. Bartlett (1937) noted, in the particular case of the hypothesis of the equality of variances for $k + 1$ normal populations, that the χ_k^2 distribution was a far better fit to the distribution of $k\Lambda/E_\theta\Lambda$ than to Λ itself. Following work by Box (1949), Lawley (1956), by ingenious and difficult cumulant calculations, "established" the folk theorem that quite generally

$$P_\theta\left[\frac{k\Lambda}{\hat{E}} \le t\right] = \chi_k^2(t) + O(n^{-2}), \tag{1.6}$$

where

$$\hat{E} = k + \frac{\hat{b}}{n} = E_\theta(\Lambda) + O_p(n^{-3/2})$$

and $\hat{b}$ is a suitable estimate for the coefficient b of n^{-1} in the expansion of $E_\theta(\Lambda)$. Departing from an asymptotic formula for the conditional density of X given an ancillary due to Barndorff-Nielsen (1986). Barndorff-Nielsen and Cox (1984) showed that (1.6) can be expected to hold quite generally and they derived formulas for estimating b in one important class of models. Efron (1985) established (for an important special case) a related result. Let

$$T = \Lambda^{1/2}\,\mathrm{sgn}(\hat{\theta}^1 - \theta^1).$$

Then

$$P_\theta[T \le t] = \Phi\left(\frac{t - \mu(\theta)}{\sigma(\theta)}\right) + O(n^{-3/2}), \tag{1.7}$$

where

$$\mu(\theta) = \frac{a_0(\theta)}{\sqrt{n}} + \frac{a_1(\theta)}{n} + O(n^{-3/2}),$$

$$\sigma^2(\theta) = 1 + \frac{c(\theta)}{n} + O(n^{-3/2}),$$

where a_0, a_1 and c are suitable functions of θ, not depending on n. As P. McCullagh pointed out to us, this result implicitly already appears in Lawley (1956) and, in fact, $a_1 = 0$. It is easy to see that, for $k = 1$, (1.7) finally implies (1.6) [with $O(n^{-2})$ replaced by $O(n^{-3/2})$] with $\hat{b}$ estimating $a_0^2(\theta) + c(\theta)$.

Our aim in this paper is:

1. To give a generalization of Efron's result to vector parameters. A closely related result appears in Barndorff-Nielsen (1986) and is again foreshadowed by Lawley (1956).
2. To apply this extension to establish the validity of Bartlett's correction for the p variate joint distribution of the Λ statistics (deviances) arising from

testing the nested hypotheses H_k: $\theta^j = \theta_0^j$, $j = 1, \ldots, k$, within H_{k-1} for $k = 1, \ldots, p$. That is, to show that, when the deviances are standardized by their asymptotic expectations to order $1/n$, their joint distribution under θ_0 differs from that of p independent identically distributed χ_1^2 variables by an error of order n^{-2}. This result is also implicit in Lawley (1956) although the calculations are purely formal. For the case of a single statistic Λ, this can be obtained in a rigorous fashion under appropriate regularity conditions from Chandra and Ghosh (1979).

3. To give Bayesian analogues of both of these results which we believe provide a key to understanding the Bartlett phenomenon. The Bayesian analogue is interesting in its own right, is fairly easy to establish and is the basic step in our arguments for aims 1 and 2.

Here is a discussion of the motivation and the structure of our Bayesian argument when we restrict to the familiar case of i.i.d. observations from a smooth parametric family. It has been proved in Chandra and Ghosh (1979) that the distributions of the likelihood ratio, as well as Wald's and Rao's score statistic, have asymptotic expansions in powers of n^{-1}, which are valid in the sense of Bickel (1974). These types of expansions have been around for a long time; see Box (1949). When viewed as formal expansions for the density $p_n(\chi^2)$ of one of these statistics, they are of the form $ce^{-\chi^2/2}(\chi^2)^{k/2-1}\{1 + \psi_1(\chi^2)n^{-1} + \cdots\}$, where the coefficients ψ are polynomials in χ^2. It is easy to check that adjustment of such a statistic through multiplication or division by a constant of the form $(1 + bn^{-1})$ will knock off the coefficient of n^{-1} in the expansion for the adjusted statistic, iff ψ_1 is linear. By examining various examples one can convince oneself that ψ_1 is not linear for Wald's or Rao's statistic. Moreover it is far from clear why ψ_1 is linear for the likelihood ratio statistic. This paper is addressed to clearing up mysteries of this kind as well as to exploring the duality between the Bayesian and the frequentist setup which, to first order, was studied extensively by Le Cam under the rubric of the Bernstein–von Mises theorem.

Our Bayesian route could be followed to produce a relatively transparent proof of linearity of ψ_1. However, since we want to do more, namely, derive the asymptotic expansion for the joint distribution of the p deviances statistics up to $O(n^{-2})$, we first note, in a similar vein, that here also the question boils down to the structure of the polynomials that appear as coefficients of powers of n^{-1} in the expansion. The relevant results for this purpose are Lemmas A2 through A4 in the Appendix. These lemmas need to be applied to the vector $T(\theta, \mathbf{X})$ of the signed square roots of the likelihood ratio statistics, defined in Section 2. That the distribution of these statistics has a valid Edgeworth expansion can be shown using Theorem 2 of Bhattacharya and Ghosh (1978). In the frequentist setup the sort of structure one needs for the polynomials is specified in the conclusion of Theorem 3. It turns out that one needs the polynomials corresponding to $n^{-1/2}$ and n^{-1} to be of degree at most 1 and 2, respectively. To prove this, one first obtains a similar result in the Bayesian setup, namely, Theorem 1, which provides an expansion for the posterior

distribution of $T(\theta, \mathbf{X})$ given $\mathbf{X}$. The likelihood factor in the posterior $\exp\{nl(\theta) - nl(\hat{\theta})\}$ is exactly the sum of squares of the components of T and so no expansion is needed. The coefficient polynomials in the asymptotic expansion arise only from the Taylor expansions of the prior density $\pi(\theta)$ around $\hat{\theta}$ and a stochastic expansion of the Jacobian of the transformation of $(\theta - \hat{\theta})$ to $T(\theta, X)$ viewed as a function of random θ. For reasons that are not hard to see, in these latter expansions the degree of the coefficient polynomial matches the power of n^{-1}; vide Lemmas 1 and 2. These facts are at the heart of the proof of Theorem 1. Theorem 1 would fail for Wald's or Rao's statistic because the likelihood factor $\exp\{nl(\theta) - nl(\hat{\theta})\}$ cannot be written as the square of either of them exactly and so an expansion of this term is called for too. Finally, Theorem 3 follows because Theorem 1 is true for a set of priors which is dense in the weak topology.

Our expansions may be used to set up Bayesian or frequentist confidence intervals; see the discussion following Corollary 1.

We propose to carry out our program without relying on the i.i.d. sampling assumption, under conditions such as those of Bickel, Götze and van Zwet (1985) which emphasize that we are, as with the original Wilks result, dealing with a phenomenon which depends only on the asymptotic stability of l and its derivatives, moderate deviation properties of $\hat{\theta}$ and related estimates and the existence of Edgeworth expansions for the distribution of T. Simple conditions implying those we give may be specified in the case of Markov and independent nonidentically distributed observations in the same way as is done in Bickel, Götze and van Zwet (1985).

A feature of our approach is that calculations are kept to a minimum so that, we believe, the phenomena are transparent. The disadvantage here is that unlike our predecessors, we do not arrive at formulae for the (estimated) coefficient $\hat{b}$ needed in the correction. It is, however, worth pointing out that, in situations which are like simple random sampling and where computing power is readily available, we can obtain $\hat{b}$ without knowing its form by applying the jackknife for bias reduction; see Efron (1982), for example. That is, we calculate Λ_{-i}, the Λ statistic for the data X_j, $j \neq i$, and put

$$\hat{b} = \sum_{i=1}^{n} (\Lambda_{-i}) - nk.$$

The paper is organized as follows. Section 2 contains the statements of the main theorems plus the necessary assumptions and notations. Section 3 contains the proofs of our results. Four simple technical lemmas are in the Appendix.

2. The main results. Since we intend to use tensor notation for arrays, we subsequently identify vector components by superscripts, for example, $\theta = (\theta^1, \ldots, \theta^p)$. For given $\theta \in \Theta$, define $\hat{\theta}_j$ as the maximum likelihood estimate of θ when $\theta^1, \ldots, \theta^j$ are fixed, i.e.,

$$l(\hat{\theta}_j) = \max\{l(\tau): \tau^1 = \theta^1, \ldots, \tau^j = \theta^j\}. \tag{2.1}$$

We shall in the sequel assume that these quantities exist and are unique but at the end of the section will sketch how this requirement can be weakened. Define $T = (T^1, \ldots, T^p)$, where

$$T^j \equiv n^{1/2}\Big[2\Big(l\big(\hat{\theta}_{j-1}\big) - l\big(\hat{\theta}_j\big)\Big)\Big]^{1/2} \operatorname{sgn}\big(\hat{\theta}^j_{j-1} - \theta^j\big). \tag{2.2}$$

Note that T is a function of θ and $\mathbf{X}$.

Let π be a prior density on Θ. Let P denote the joint distribution of $(\theta, \mathbf{X})$ and $P(\cdot|\mathbf{X})$ the conditional (posterior) probability distribution of $(\theta, \mathbf{X})$ given $\mathbf{X}$. Let $r = n^{-1/2}$ and consider the posterior density of $r^{-1}(\theta - \hat{\theta})$ given by

$$\pi(h|\mathbf{x}) \equiv \exp\{l(\hat{\theta} + rh) - l(\hat{\theta})\}\pi(\hat{\theta} + rh)/N(\mathbf{x}),$$

where

$$N(\mathbf{X}) = \int \exp\{l(\hat{\theta} + rh) - l(\hat{\theta})\}\pi(\hat{\theta} + rh)\, dh. \tag{2.3}$$

Let

$$\phi(t) = (2\pi)^{-p/2} \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{p} (t^i)^2\right\}$$

be the standard p variate normal density. Let $\pi_T(t|\mathbf{X})$ denote the posterior density of T [which exists under our assumptions with probability $1 - O(r^{m+1})$].

NOTATION. We postulate $m + 3$ continuous derivatives for $l(\theta)$, $\pi(\theta)$ and write $l_{i_1 \ldots i_k}$ for $\partial^k l/\partial\theta^{i_1} \cdots \partial\theta^{i_k}$, etc. Following tensor notation, we indicate arrays by their elements. Thus l^i is a vector, l_{ij} a matrix, etc. We also follow the Einstein convention of summing over a subscript which is repeated in a superscript, e.g., $l_{ij}l^i = \sum_i l_{ij}l^i$. Occasionally we denote a vector array by symbols like ν_i, so that $\nu_i t^i$ stands for $\sum_i \nu_i t^i$.

Here are the main results stated under regularity conditions which appear at the end of the section.

THEOREM 1. *If B_m holds, then*

$$E_P \int |\pi_T(t|\mathbf{X}) - \pi_m(t, \mathbf{X})|\, dt = O(r^{m+1}), \tag{2.4}$$

where

$$\pi_m(t, \mathbf{X}) = \phi(t)(1 + P_m(r, \mathbf{X}, \pi) + Q_m(rt, \mathbf{X}, \pi))1(\mathbf{X} \in S),$$

P_m is a polynomial in r of degree m, Q_m is a polynomial in rt of degree m [both without constant terms and with coefficients which are rational functions of $l_{h_1 \ldots h_k}(\hat{\theta})$] and $\pi_{h_1 \ldots h_k}(\hat{\theta})/\pi(\hat{\theta})$ for $1 \le k \le n + 2$ and $P[X \notin S] = O(r^{m+1})$ where S is given in Section 3. 1(A), as usual, denotes the indicator of A.

Write

$$P_m(r, \mathbf{X}, \pi) = \sum_{k=1}^{m} P_{mk}(\mathbf{X}, \pi) r^k,$$

$$Q_m(u, \mathbf{X}, \pi) = \sum_{k=1}^{m} Q_{mb_1 \cdots b_k}(\mathbf{X}, \pi) u^{b_1} \cdots u^{b_k}$$

and note that P_m, Q_m and S depend on n.

NOTE 1. It is necessary to keep the indicator of S in π_m since the coefficients $P_{mk}, Q_{mb_1 \cdots b_k}$ need not be bounded outside S.

The proof of Theorem 1 actually also yields that if $X \in S$, i.e., with probability $1 - O(r^{m+1})$, the random quantity

$$\int |\pi_T(t|\mathbf{X}) - \pi_m(t|\mathbf{X})| \, dt$$

is $O(r^{m-1})$.

NOTE 2. Since P_{mk} and $Q_{mb_1 \cdots b_k}$ depend on r they are not uniquely defined. Since

$$E\left|\int \pi_m(t, \mathbf{X}) \, dt - 1\right| = O(r^{m+1}). \tag{2.4'}$$

It is easy to see that we can always take $P_{m0} = Q_{m0} = 0$ and suppose all P_{mk} for k odd to be zero. For example, suppose we are given a set of $P_{mk}^{(1)}$ and associated Q_m. Note that

$$P_{m1}^{(1)} r + \int Q_{m1} r t \phi(t) \, dt = O(r^2) \quad \text{if } m \geq 1.$$

Therefore, $P_{m1}^{(1)} = O(r)$. Hence we can define the following set $P_{mk}^{(2)}$ satisfying (2.4): $P_{m0}^{(2)} = 0$, $P_{m2}^{(2)} = P_{m1}^{(1)} r + P_{m2}^{(1)} + P_{m3}^{(1)} r$, $P_{mk}^{(2)} = 0$ for k odd and $P_{mk}^{(2)} = P_{mk}^{(1)} + r P_{m(k+1)}^{(1)}$ for k even and greater than or equal to 4.

NOTE 3. Note that (2.4') for $m = 2, 3$ implies

$$E\left|\int \pi_2(t, X) \, dt - 1\right| = E\left| P_{22} r^2 - Q_{2ij} \delta^{ij} r^2 \right| = O(r^3).$$

In view of Notes 1 and 2 and the above relation we deduce, putting $m = 1, 2$ in (2.4), that with probability $1 - O(r^2)$ and $1 - O(r^3)$, respectively, the posterior distribution of T is $N_p(rQ_{1i}, J)$ with error $O(r^2)$ and $N_p(rQ_{2i}, J + r^2(2Q_{2ij} - Q_{2i}Q_{2j}))$ with error $O(r^3)$, where $N_p(\mu, \Sigma)$ is the p variate normal distribution with mean μ and dispersion matrix Σ and J is the $p \times p$ identity matrix. These are the multivariate Bayesian analogues of Efron's (1985) result.

NOTE 4. The relation (2.4') for $m = 3$ implies as above that

$$E\left| P_{32} r^2 - Q_{3ij} \delta^{ij} r^2 \right| = O(r^4)$$

and hence that π_3 may be written as

$$rQ_{3i}t^i + r^2Q_{3ij}(t^it^j - \delta^{ij}) + r^3Q_{3ijk}t^it^jt^k + O(r^4),$$

which has the structure of $g(t)$ of Lemma A2 up to $O(r^4)$. This fact will be used in the proof of Theorem 2.

Let $c_k(\cdot)$ denote the χ^2_k density,

$$D^j \equiv (T^j)^2 = 2n\left(l\left(\hat{\theta}_{j\ 1}\right) - l\left(\hat{\theta}_j\right)\right)$$

the deviance and

$$\tilde{D}^j = D^j/\left(1 + 2r^2Q_{2jj}\right)$$

the standardized (Bartlett corrected) deviance. If π_D and $\pi_{\tilde{D}}$ are the corresponding posterior densities of these vectors $D = (D^1, \ldots, D^p)$ and $\tilde{D} = (\tilde{D}^1, \ldots, \tilde{D}^p)$, then one has the following result.

THEOREM 2. *Under* B_1,

$$E_P\left\{\int\left|\pi_D(u|\mathbf{X}) - \prod_{j=1}^{p} c_1(u^j)\right| dt\, 1(\mathbf{X} \in S)\right\} = O(n^{-1}), \tag{2.5}$$

while under B_3,

$$E_P\left\{\int\left|\pi_{\tilde{D}}(u|\mathbf{X}) - \prod_{j=1}^{p} c_1(u^j)\right| du\, 1(\mathbf{X} \in S)\right\} = O(n^{-2}). \tag{2.6}$$

In fact (vide Note 1), with probability $1 - O(n^{-1})$ and error $O(n^{-1})$ the posterior distribution of D is that of p independent χ^2_1, while for $\tilde{D}$ the same claim holds with probability $1 - O(n^{-2})$ and error $O(n^{-2})$.

From this we deduce:

COROLLARY 1. (a) *Under* B_1, *if* π_Λ *is the posterior distribution of* Λ *given by* (1.4),

$$E_P\left\{\int|\pi_\Lambda(u|\mathbf{X}) - c_k(u)|\, du\, 1(\mathbf{X} \in S)\right\} = O(n^{-1}). \tag{2.7}$$

(b) *Let* $\tilde{\Lambda} = \Lambda/(1 + 2r^2k^{-1}\Sigma_{j=1}^{k} Q_{2jj})$. *Then, under* B_3,

$$E_P\left\{\int|\pi_{\tilde{\Lambda}}(u|\mathbf{X}) - c_k(u)|\, du\, 1(\mathbf{X} \in S)\right\} = O(n^{-2}). \tag{2.8}$$

So (2.7) says that the posterior distribution of Λ is χ^2_k with error $O(n^{-1})$ while (2.8) is the Bayesian analogue of the Bartlett phenomenon. The posterior distribution of the Bartlett standardized statistic $\tilde{\Lambda}$ is χ^2_k with error $O(n^{-2})$.

These results can in principle be used to set Bayesian posterior confidence regions for θ to order n^{-1}, n^{-2} in a variety of ways. For instance, $\{\theta$: $\Lambda \le \chi_p(1-\alpha)\}$ where χ_p is the $1-\alpha$ percentile of χ^2_p and $\Lambda = 2(l(\hat{\theta}) - l(\hat{\theta}))$ has posterior probability $1-\alpha$ with error $O(n^{-1})$, while $\{\theta$: $\tilde{\Lambda} \le \chi_p(1-\alpha)\}$

has posterior probability $1 - \alpha$ with error $O(n^{-2})$. Of course regions could be based on other functions of D_j and $\tilde{D}_j$, for instance, on $\max_j D_j$ or $\max_j \tilde{D}_j$. They could also be used in investigating the old question of what choices of model and prior lead to posterior probability regions which are also frequentist regions with error $O(n^{-2})$; see, for example, Stein (1985) and Welch and Peers (1963). However, more detailed computation of the Q_j than we provide seems necessary for this endeavor.

We use these results only in establishing the corresponding result in the frequentist case.

THEOREM 3. *Suppose that F_m holds and the density of T, $p_T(t|\theta)$, admits an Edgeworth expansion such that if $i^2 = -1$,*

$$(2.9) \qquad \left| \int e^{i\nu_j t^j} \left[p_T(t|\theta) - \phi(t) \left\{ 1 + \sum_{k=1}^{m} r^k R_k(t,\theta) \right\} \right] dt \right| = O(r^{m+1})$$

uniformly in compact sets of θ and ν, where the $R_k(\cdot,\theta)$ are continuous in θ and polynomials in t, independent of r. Then, the R_k are of at most degree k in t.

As in Notes 2 and 3, it is clear that (2.9) implies, on taking $\nu = 0$, that $R_1(t,\theta) = R_{1j}t^j$ and $R_2(t,\theta) = R_{2ij}(t^i t^j - \delta^{ij}) + R_{2i}t^i$, where δ^{ij} is the Kronecker delta. In the following we shall need a condition analogous to (2.9), namely,

$$(2.9') \qquad \left| \int e^{i\nu_j (t^j)^2} \left[p_T(t|\theta) - \phi(t) \left\{ 1 + \sum_{k=1}^{m} r^k R_k(t,\theta) \right\} \right] dt \right| = O(r^{m+1})$$

uniformly in compact sets of θ and all ν. We deduce our generalization of Efron's result.

COROLLARY 2. *If $m = 1$, the characteristic function of p_T differs from that of $N(rR_{1j}, J)$ by $O(r^2)$ and if $m = 2$, from $\mathbf{N}(rR_{1j}, J + r^2(2R_{2j} - R_{1i}R_{1j}))$ by $O(r^3)$.*

THEOREM 4. *If the assumptions of Theorem 3 and (2.9') hold for $m = 1$, then, uniformly in ν,*

$$(2.10) \qquad \int e^{i\nu u} \left[p_D(u|\theta) - \prod_{j=1}^{p} c_1(u^j) \right] du = O(n^{-1}),$$

i.e., the approximation $\prod_{j=1}^{p} c_1(u^j)$ is good to order n^{-1}.

Further, let

$$\tilde{D}^j = D^j / \left(1 + 2r^2 R_{2jj}\right).$$

If (2.9), (2.9') and F_m hold for $m = 3$, then uniformly in ν,

$$(2.11) \qquad \int e^{i\nu_j u^j} \left[p_{\tilde{D}}(u|\theta) - \prod_{j=1}^{p} c_1(u^j) \right] du = O(n^{-2}).$$

COROLLARY 3. *Under the conditions of Theorem* 4, *uniformly in* ν,

$$\int e^{ivu}\left[p_{\Lambda}(u|\theta) - c_k(u)\right] du = O(n^{-1}), \tag{2.12}$$

$$\int e^{ivu}\left[p_{\tilde{\Lambda}}(u|\theta) - c_k(u)\right] du = O(n^{-2}). \tag{2.13}$$

It turns out that $T^i = r^{-1}(\hat{\eta}^i - \eta^i) + O(r)$ [see (3.6) and (3.19)] and $r^{-1}(\hat{\eta}^i - \eta^i)$ is up to $O(r)$ a linear function of the first derivatives of the log likelihood evaluated at θ. In fact it is possible to stochastically expand T in terms of the derivatives of the log likelihood evaluated at θ, with a leading linear term. In the i.i.d. case if enough moments are finite, we can talk of a formal Edgeworth expansion for the density or distribution function of T and under the same assumptions the rigorous expansion of the characteristic function of T that we require is valid; vide the introduction in Bhattacharya and Ghosh (1978). This is all that one needs to justify the Bartlett correction and the related results as given in Theorem 4. If one wants these results to be valid for the distribution function in the sense of Bickel (1974), it is enough to assume that the Edgeworth expansion for the density of T is valid in the L_1 sense. This assumption may be verified via Theorem 2(a) of Bhattacharya and Ghosh (1978) if the derivatives of the log likelihood appearing in the stochastic expansion for T up to $o_p(n^{-3/2})$ have an absolutely continuous joint distribution. Actually, instead of absolute continuity, it is enough to assume Cramer's condition [vide condition C of Bhattacharya and Ghosh (1978)] and apply their Theorem 2(b) instead of Theorem 2(a).

We note again that a form of Theorem 4 appeared in Barndorff-Nielsen (1986) [with error $O(n^{-3/2})$]. Barndorff-Nielsen's results focus on conditional inference given asymptotic ancillary statistics. His work implicitly requires conditions for the validity of saddlepoint expansions for the conditional density. These in turn imply but are not necessary for the validity of Edgeworth expansions for the conditional density. The Edgeworth expansions may be used in conjunction with our "Bayesian" result to derive the appropriate analogues of Theorem 4. We believe our Bayesian route makes matters easier and more transparent. The assumptions below may appear rather strong but, as indicated in the remarks, they hold quite generally. Moreover, they are quite natural if one is to develop a rigorous, rather than a formal, argument.

Suppose we estimate the correction factor and adjust the likelihood ratio statistic in (1.6). If in Corollary 3 we replace $\tilde{\Lambda}$ by $k\Lambda/(k + \hat{b}/n)$ then the conclusion of Corollary 3 holds under suitable regularity conditions. This fact was first noted by Barndorff-Nielsen and Hall (1988). The most brutal condition is to suppose that

$$\hat{b} = b(\theta) + rc_i t^i + \Delta(\theta), \tag{2.14}$$

where

$$E_\theta|\Delta(\theta)| = O(r^2).$$

Of course (2.14) is motivated by a stochastic expansion such as

$$\hat{b} \equiv b(\hat{\theta}) = b(\theta) + d_i(\hat{\theta}^i - \theta^i) + O_p(r^2) \tag{2.15}$$

and the expansion

$$\hat{\theta}^i - \theta^i = r\hat{D}_{ij}T^j + O_p(r^2)$$

for a suitable $\hat{D}_{ij}$; see Lemma 2. To show that (2.14) and the assumptions of Corollary 3 are enough for this result we need only note that the difference between the Fourier transforms of $\tilde{\Lambda}$ and $k\Lambda/(k + \hat{b}/n)$ at ν can be written [with an appropriate constant $M(\theta)$] as

$$M(\theta)\int \exp\left[\left(-\tfrac{1}{2}\sum_{i=1}^{p}[t^i]^2\right) + i\nu\Sigma[t^i]^2\right]\left[\Sigma[t^i]^2(c_i t^i)\right]r^3\,dt + O(r^4)$$

uniformly on compact ν subsets. The integral vanishes by symmetry.

Condition (2.14) is too brutal but can readily be replaced by the possibility of further expansion of (2.15) and large deviation estimates for $\hat{\theta} - \theta$. Alternatively, we can simply suppose that the Edgeworth expansion of $k\Lambda(k + \hat{b}r^2)^{-1}$ agrees with that of $\tilde{\Lambda}(1 - (k + b(\theta)r^2)^{-1}r^2c_iT^i)$ with error of order r^3. This kind of replacement can be proved in a standard fashion under the usual protocols for asymptotic expansions of maximum likelihood estimates; see Pfanzagl (1974), for example.

We postulate nonrandom arrays λ_i, λ_{ij}, etc. and write,

$$l_{i_1\cdots i_k}(\theta) = \lambda_{i_1\cdots i_k}(\theta) + \Delta_{i_1\cdots i_k}(\theta).$$

Here are our conditions. Let $|\cdot|$ denote the l_1 norm on R^p. For all $0 < M < \infty$ and some $0 < \delta < 1$, $\varepsilon_n \downarrow 0$.

B_m: (i) $P[|\hat{\theta} - \theta| \geq Mr^{1-\delta}] = O(r^{m+1})$.

(ii) $P[|\hat{\theta} - \theta| \leq Mr^{m+2}] = O(r^{m+1})$.

Let

$$A = \left\{\mathbf{x}\text{: for all } j, \left\{\theta\text{: } |\hat{\theta}(\mathbf{x}) - \theta| \leq M_1 r^{1-\delta}\right\} \subset \left\{\theta\text{: } |\hat{\theta}_j(\mathbf{x},\theta) - \hat{\theta}(\mathbf{x})| \leq M_2 r^{1-\delta}\right\}\right\}.$$

For all $0 < M_1 < \infty$, there exists $0 < M_2 < \infty$ such that:

(iii) $P[\mathbf{X} \notin A] = O(r^{m+1})$.

(iv) $P[\sup\{|\Delta_{i_1\cdots i_k}(\hat{\theta} + r\nu)|\text{: } |\nu| \leq Mr^{1-\delta}\} \geq \varepsilon_n] = O(r^{m+1})$, $1 \leq k \leq m + 3$.

(v) The maps $\theta \to \lambda_{i_1\cdots i_k}(\theta)$ are continuous, $1 \leq k \leq m$.

(vi) The matrix $\|-\lambda_{ij}(\theta)\|$ is positive definite for all θ.

(vii) (a) π vanishes off a compact $K \subset \Theta$. (b) $P[\sup\{|\pi_{i_1\cdots i_{m+2}}(\hat{\theta} + r\nu)|/\pi(\hat{\theta})\text{: } |\nu| \leq Mr^{-\delta}\} \geq r^{-\delta}] = O(r^{m+1})$.

F_m: Uniformly on compacts in θ:

(i) $P_\theta[|\hat{\theta} - \theta| \geq Mr^{1-\delta}] = O(r^{m+1})$.

(ii) $P_\theta[|\hat{\theta} - \theta| \leq Mr^{m+2}] = O(r^{m+1})$.

(iii) $P_\theta[\mathbf{X} \notin A] = O(r^{m+1})$ for A defined in B_m.

(iv) $P_\theta[\sup\{|\Delta_{i_1\cdots i_k}(\theta + r\nu)|\text{: } |\nu| \leq Mr^{1-\delta}\} \geq \varepsilon_n] = O(r^{m+1})$, for $1 \leq k \leq m + 3$.

(v) Condition (v) of B_m.

(vi) Condition (vi) of B_m.

REMARKS. (a) We give a qualitative discussion of the "Bayesian" conditions B_m. The frequentist conditions F_m can be viewed in an analogous fashion.

(i) Variations of the mle $\hat{\theta}$ from θ of order $n^{-1/2(1-\delta)}$ occur with very small probability. Thus we can safely think about Taylor expanding $l(\theta)$ and $l(\hat{\theta}_j(\theta))$ around $\hat{\theta}$.

(ii) This condition says that $r^{-1}(\hat{\theta} - \theta)$ has approximately a bounded density near 0. It is needed to ensure that the map $\theta - \hat{\theta} \to \mathbf{T}(\theta, \mathbf{x})$ is 1–1 and otherwise well behaved with high probability.

(iii) This condition assumes that both $\hat{\theta}$ and $\hat{\theta}_j$ are close to θ and each other
simultaneously. It is needed for expansions of $l(\hat{\theta}_j(\theta))$.

(iv) The coefficients of the Taylor expansion differ little from constants, or more specifically, $l(\theta)$ and its derivatives behave like averages of i.i.d. variables.

(v) Smoothness conditions needed to permit replacement of quantities such as $\lambda_{i_1 \cdots i_k}(\hat{\theta}_j(\theta))$ appearing as approximations to coefficients in the Taylor expan-
sion of $l(\hat{\theta}_j(\theta))$ by $\lambda_{i_1 \cdots i_k}(\hat{\theta})$.

(vi) Nonsingularity of the information matrix is necessary even for the statement of the Bernstein–von Mises theorem.

(vii) We need to expand $\log \pi(\theta)$ around $\hat{\theta}$. Condition (a) is useful for technical reasons, while (b) is needed to control $\log \pi$ and its derivatives near the boundary of K where $\log \pi \to -\infty$.

(b) The validity of F_m and B_m other than (ii) and (iii) has been checked for independent nonidentically distributed and Markov dependent observations in Bickel, Götze and van Zwet (1985). In particular these conditions hold for exponential families in the i.i.d. case. They also hold in many examples for such families in the independent nonidentically distributed case, e.g., in regression and GLIM models. Another example is the class of aperiodic irreducible finite state Markov chains with stationary completely unknown transition matrix.

(c) Condition B_m(ii) in fact follows from the other B_m conditions since they guarantee an Edgeworth expansion for $\pi(h|\mathbf{X})$. An Edgeworth expansion uniform on θ compacts for the distribution of $r^{-1}(\hat{\theta} - \theta)$ implies F_m(i) and (ii). Condition F_m or B_m(iii) holds if the log likelihood is convex.

(d) The conditions on existence of the estimate $\hat{\theta}_j$ can be replaced by requiring the existence of a preliminary estimate $\tilde{\theta}$ with appropriate moderate deviation properties and then redefining the $\hat{\theta}_j$ as the result of $m + 1$ iterations of the Newton–Raphson method applied to the appropriate likelihood equations. See Theorem 4 of Bickel, Götze and van Zwet (1985).

(e) In the situation of (d), suppose that F_m(iv)–(vi) hold and that, uniformly on θ compacts, for all $0 < M < \infty$,

$$P_\theta\left[|\tilde{\theta} - \theta| \geq Mr^{1-\delta}\right] = O(r^{m+1}),$$
$$P_\theta\left[|\tilde{\theta} - \theta| \leq Mr^{m-2}\right] = O(r^{m+1}). \tag{2.16}$$

Let

$$A^* = \left\{\mathbf{x}\text{: for all } j, \left\{\theta\colon |\tilde{\theta} - \theta| < M_1 r^{1-\delta}\right\} \subset \left\{\theta\colon \left|\hat{\theta}_j - \tilde{\theta}\right| < M_2 r^{1-\delta}\right\}\right\}.$$

Then uniformly on θ compacts,

$$P_\theta[\mathbf{X} \in A^*] = O(r^{m+1}).$$

If we redefine the set B of Section 3 so that B(ii) is replaced by

$$|\hat{\theta}^b - \theta^b| > M^* r^{m+2}, \qquad |\tilde{\theta}^b - \theta^b| < r^{1-\delta},$$

then the proof of Theorems 4 and 5 goes through.

3. Proofs. We need to analyze $\pi_T(t|\mathbf{X})$ where we assume that $\mathbf{X}$ belongs to
a set S on which the map $h \to T(\hat{\theta} + rh, \mathbf{X})$, $|h| < Mr^{-\delta}$, is invertible with nonvanishing Jacobian and the matrix $\| -l_{ij}(\hat{\theta})\| = \hat{C}$ is positive definite. We explain the transformation in more detail and give S below. Let $\hat{D}$ be the unique lower triangular matrix with positive diagonal such that

$$\hat{D}\hat{D}^T = \hat{C} \tag{3.1}$$

and

$$L(\eta) = l(\hat{D}^{-1}\eta). \tag{3.2}$$

If $\|l_{ij}(\hat{\theta})\|$ is the Hessian of l at $\hat{\theta}$ and $\hat{\eta} = \hat{D}\hat{\theta}$, then in the usual notation,

$$-L_{ij}(\hat{\eta}) = J, \tag{3.3}$$

the $p \times p$ identity. This in the Bayesian domain corresponds to standardizing the Fisher information at θ to be J as is done in the corresponding frequentist calculations. Further define $\hat{\eta}_j$ by

$$L(\hat{\eta}_j) = \max\{L(\gamma)\colon \gamma^1 = \eta^1, \ldots, \gamma^j = \eta^j\} \tag{3.4}$$

and

$$\tilde{T}^i(\eta) = r\left(2(L(\hat{\eta}_{i-1}) - L(\hat{\eta}_i))\right)^{1/2} \operatorname{sgn}\left(\hat{\eta}^i_{i-1} - \eta^i\right). \tag{3.5}$$

It is easy to verify that

$$T(\hat{\theta} + rh) = \tilde{T}(\hat{\eta} + r\hat{D}h). \tag{3.6}$$

Now $\hat{D}r^{-1}(\theta - \hat{\theta})$ has posterior density

$$\pi(\hat{D}^{-1}h|\mathbf{X})\left|\det(\hat{D})\right|^{-1} \tag{3.7}$$

and hence

$$\pi_T(t|\mathbf{X}) = \exp\left(-\tfrac{1}{2}\sum_{i=1}^{p}(t^i)^2\right)\pi\left(\hat{D}^{-1}(\hat{\eta} + rh(t))\right)\det\left\|h^i_j(t)\right\|/M(\mathbf{X}), \tag{3.8}$$

where $h(t)$ is defined by

$$\tilde{T}(\hat{\eta} - rh(t)) = t \tag{3.9}$$

and

$$h_j^i(t) = \frac{\partial h^i}{\partial t_j}(t),$$

$$M(\mathbf{X}) = \int \exp\left(-\tfrac{1}{2}\sum_{i=1}^{p}(t^i)^2\right)\pi(\hat{D}^{-1}(\hat{\eta} + rh(t)))\det\|h_j^i(t)\|.$$

For fixed $\mathbf{X}$, let R_X be the image of $\{h: |h| < Mr^{-\delta}\}$ under the map $h \to T(\hat{\theta} + rh, \mathbf{X})$. From (3.8) it is clear that our task in proving Theorem 1 is to exhibit the set S such that, for $t \in R_X$, h is uniquely defined by (3.9) and such that

$$h(t) = t + rP(t, \mathbf{X}) + O(r^{m+1}), \tag{3.10}$$

$$h_j^i(t) = \delta_{ij} + rP_{ij}(t, \mathbf{X}) + O(r^{m+1}), \tag{3.11}$$

where P and P_{ij} are polynomials in t, and to identify the order of the polynomials. Here $O(r^{m-1})$ means that the remainder is bounded on S by Mr^{m+1} for a generic constant M independent of n.

We define B as the set where

(i) $\sup\{|\pi_{i_1 \cdots i_{m+2}}(\hat{\theta} + r\nu)|/\pi(\hat{\theta}): |\nu| \le Mr^{-\delta}\} \le r^{-\delta}$.

(ii) $M^* r^{m-2} < |\hat{\theta}^b - \theta^b| < r^{1-\delta}, \quad 1 \le b \le p$.

(iii) $\sup\{|\Delta_{i_1 \cdots i_s}(\hat{\theta} + r\nu)|: |\nu| \le Mr^{-\delta}\} \le \varepsilon_n$.

Note that, by B_m,

(a) $P[(r^{-1}(\hat{\theta} - \theta), \mathbf{X}) \in B^c] = O(r^{m+1})$.

(b) The $\mathbf{x}$ sections of B intersect each quadrant in an open convex set since $|\cdot|$ is the l_1 norm.

(c) There exists a generic constant $C > 0$ such that on B,

$$\sup\{|l_{i_1 \cdots i_k}(\hat{\theta} + rh)|: |h| \le Mr^{-\delta}\} \le C.$$

(d) $C^{-1} \le \lambda \le \bar{\lambda} \le C$ where $\lambda, \bar{\lambda}$ are the minimal and maximal eigenvalues of $\|-l_{ij}(\hat{\theta})\|$.

(e) $|\hat{\theta}_j - \hat{\theta}_{j-1}| \le M_2 r^{1-\delta}$, $|\hat{\theta}_j - \hat{\theta}| \le M_1 r^{1-\delta}$.

We let $\tilde{S}$ be the image of B under the map $(h, x) \to (T(\theta(x) + rh, x), x)$ and S be just the projection of $\tilde{S}$ on the $\mathbf{x}$ axis, i.e., the set of all $\mathbf{x}$ satisfying (i) and (iii) above.

CONVENTION. Expressions such as $\hat{\eta}_i(\eta)$ are calculated at $\eta = \hat{\eta} + rh$.

LEMMA 1. *On B, for $j \ge i + 1$,*

$$\hat{\eta}_i^j = \hat{\eta}^j + \sum_{k=2}^{m+1} N_{b_1 \cdots b_k}^{ij} r^k h^{b_1} \cdots h^{b_k} + O(r^{m+1}), \tag{3.12}$$

where $N_{i_1 \cdots i_k}$ are polynomials in the derivatives $L_{i_1 \cdots i_t}$ of L (evaluated at $\hat{\eta}$)

with $t \le k$ and $h = r^{-1}(\eta - \hat\eta)$ with no constant term. Let $d = \hat\eta^i_{i-1} - \eta^i$. Then

$$\hat\eta^j_{i-1} - \hat\eta^j_i = \sum_{k=1}^{m+1} M^{ij}_k d^k + O(|d|^{m+2}), \tag{3.13}$$

where M^{ij}_k are polynomials in $L_{i_1 \cdots i_k}$ and rh which vanish at $h = 0$.

PROOF. Write L_{ab}, etc., for derivatives of L evaluated at $\hat\eta$. For $j \ge i+1$,

$$\begin{aligned} 0 = L_j(\hat\eta_i) - L_j(\hat\eta) = L_{jb}\big(\hat\eta^b_i - \hat\eta^b\big) + \cdots + \frac{1}{(m+1)!} L_{j,b_1 \cdots b_{m+1}} \\ \times \prod_{k=1}^{m+1} \big(\hat\eta^{b_k}_i - \hat\eta^{b_k}\big) + O(r^{m+1}). \end{aligned} \tag{3.14}$$

To see this, note first that $\hat\eta_i = \hat D \hat\theta_i$ and hence, in view of (e), $|\hat\eta_i - \hat\eta| \le M_3 r^{1-\delta}$. Therefore, applying (c) and (d), again the relevant derivatives of order up to $m+2$ of L at $\hat\eta$ are bounded and (3.14) follows. Note that by (3.3), $L_{ab} = -\delta_{ab}$ and that

$$\hat\eta^b_i - \hat\eta^b = -rh^b \quad \text{for } b \le i.$$

So we can rewrite (3.14) in the form

$$\delta_{jb} u^b = P_j(u, rh) + O(r^{m+1}), \qquad j \ge i+1, \tag{3.15}$$

where $u^b = \hat\eta^b_{i-1} - \hat\eta^b_i$ and P_j is a polynomial of degree $(m+1)$ in u and rh with no term of combined degree less than 2 and bounded coefficients which are polynomials in the $L_{i_1 \cdots i_t}$.

Claim (3.12) follows from a standard Lagrange inversion argument. For (3.13) write, for $j \ge i+1$,

$$0 = L_j(\hat\eta_i) - L_j(\hat\eta_{i-1}) = -L_{jh}(\hat\eta^*) e^b, \tag{3.16}$$

where $\hat\eta^*$ is an intermediate value and $e^b = \hat\eta^b_{i-1} - \hat\eta^b_i$.

Note that

$$e^b = 0, \qquad b \le i-1, \qquad e^i = d \tag{3.17}$$

and

$$L_{jb}(\hat\eta^*) = -\delta_{jb} - O(r),$$

so that (3.16) yields, for $j \ge i+1$,

$$\big|\hat\eta^j_{i-1} - \hat\eta^j\big| = O(r)|d|. \tag{3.18}$$

Expand further to get

$$\begin{aligned} L_{jb}(\hat\eta_{i-1}) e^b + \cdots + \frac{1}{(m+1)!} L_{jb_1 \cdots b_{m+1}}(\hat\eta_{i-1}) e^{b_1} \cdots e^{b_m} \\ + O(|d|^{m+2}) = 0. \end{aligned} \tag{3.19}$$

Rewrite (3.19) in the form

$$A_{jb}e^b + A_{jb_1b_2}e^{b_1}e^{b_2} + A_{jb_1\cdots b_{m+1}}e^{b_1}\cdots e^{b_{m+1}} = a_1 d + \cdots + a_{m+1}d^{m+1} + O(d^{m+1}),$$

where the indices $b, b_1, \ldots, b_m$ range from $i+1$ to p,

$$A_{jb_1\cdots b_k} = \frac{L_{jb_1\cdots b_k}}{k!}(\hat{\eta}_{i-1})$$

and the a_i are polynomials in the $L_{jb_1\cdots b_k}(\hat{\eta}_{i-1})$ and the e^b. Expand $A_{jb_1\cdots b_k}$ around $\hat{\eta}$ to $m+1-k$ terms and use (3.12) to conclude that with remainder $O(r^{m+1})$, all the $A_{jb_1\cdots b_k}$ are polynomials in $L_{jb_1\cdots b_t}$ and rh. Finally note that, for $b \geq i+1$,

$$e^b = \hat{\eta}^b_{i-1} - \hat{\eta}^b_i = (\hat{\eta}^b_{i-1} - \hat{\eta}^b) - (\hat{\eta}^b_i - \hat{\eta}^b)$$

can by (3.12) itself be written as a polynomial of rh and $L_{jb_1\cdots b_t}$ so that the a_j are also, up to order $m+1$, polynomials in rh and $L_{jb_1\cdots b_t}$, for $t \leq m+1$. The lemma follows. □

LEMMA 2. *On B*

$$\hat{T}^i(\hat{\eta} + rh) = h^i + r^{-1}Q^i(rh) + O(r^{m+1}),$$

where Q is a polynomial of degree $m+1$ in rh with no constant or linear term and coefficients which are polynomials in $L_{b_1\cdots b_k}$, $k \leq m+2$.

PROOF. By definition

$$\hat{T}^i(\hat{\eta} + rh) = r^{-1}\left[-\sum_{k=1}^{m+2}\frac{2}{k!}L_{b_1\cdots b_k}(\hat{\eta}_{i-1})\prod_{t=1}^{k}(\hat{\eta}^{b_t}_{i-1} - \hat{\eta}^{b_t}_i) + O(|\hat{\eta}_{i-1} - \hat{\eta}_i|^{m+3})\right]^{1/2}\operatorname{sgn}(\hat{\eta}^i_{i-1} - \eta^i). \tag{3.20}$$

Note that $L_b(\hat{\eta}_{i-1}) = 0$, $b \geq i$, and $\hat{\eta}^b_{i-1} = \hat{\eta}^b_i$, $b \leq i-1$, so that the first term in the sum vanishes. Expand the coefficients around $\hat{\eta}$ and use (3.18) and (3.13) to get

$$\hat{T}^i(\hat{\eta} + rh) = r^{-1}\left(d + \sum_{k=2}^{m+2} c_k d^k + O(r^{-1}|d|^{m+2})\right), \tag{3.21}$$

where the c_k are polynomials in rh. Now substitute for d from (3.12),

$$d = rh^i + \sum_{k=2}^{m+1} N^i_{b_1\cdots b_k} r^k h^{b_1}\cdots h^{b_k} + O(r^{m+1}), \tag{3.22}$$

and the lemma follows. □

LEMMA 3. (i) *If O_i, $i = 1, \ldots, 2^p$, are the quadrants of R^p, then $\hat{T}(\hat{\eta} + rh)$ maps $O_i \cap B_{\mathbf{x}}$ into O_i for all i.*

(ii) *$\hat{T}$ is continuously differentiable on $O_k \cap B_{\mathbf{x}}$ for $1 \leq k \leq 2^p$. Let*

$$\dot{T}^i_j = \frac{\partial \hat{T}^i}{\partial h_j}.$$

Then $\tilde{T}_j^i$ is lower triangular and

$$\tilde{T}_i^i = 1 + P^i(rh) + O(r^{m-1}), \tag{3.23}$$

where P^i is a polynomial of degree $m+1$ with no constant term and coefficients in $L_{b_1 \cdots b_k}$, $k \le m+2$.

(iii) *$\tilde{T}$ is* 1–1.

PROOF. (i) We need to show that on B,

$$\operatorname{sgn}\left(\hat{\eta}_{i-1}^i - \eta^i\right) = \operatorname{sgn} h, \qquad i = 1, \ldots, p. \tag{3.24}$$

By (3.12) on B,

$$\hat{\eta}_{i-1}^i - \eta^i = rh^i(1 + rM_1(h)) + r^{m+2}M_2(h),$$

where M_1 is a polynomial in h with bounded coefficients and $|M_2(h)|$ is bounded by M_2 for all $(\mathbf{x}, h) \in B$. But $(\mathbf{x}, h) \in B \Rightarrow aM^*r^{m+1} < |h^i| < a^{-1}r^{-\delta}$, where a is positive constant depending only on the constant C of (d).

Choose M^* so that

$$aM^* > M_2. \tag{3.25}$$

The relation (3.24) follows from

$$\hat{\eta}_{i-1}^i\left(\hat{\eta} + aM^*r^{m+2}\right) - \eta^i > (aM^* - M_2)r^{m+2} + O(r^{m+3}) > 0 \tag{3.26}$$

and

$$\frac{d}{dh^i}\left\{h^i(1 + rM_1(h))\right\} = 1 + O(r).$$

(ii) It is easy to see that $\tilde{T}(\hat{\eta} + rh)$ is continuously differentiable on B with derivatives

$$\tilde{T}_j^i = |\tilde{T}^i|^{-1}\left(L_k(\hat{\eta}_{i-1})\frac{\partial \hat{\eta}_{i-1}^k}{\partial h^j} - L_k(\hat{\eta}_i)\frac{\partial \hat{\eta}_i^k}{\partial h^j}\right).$$

Note that,

$$\frac{\partial \hat{\eta}_i^a}{\partial \eta^b} = \begin{cases} 0, & a, b \ge i, \\ \delta_{ab}, & a \le i-1, \end{cases}$$

and $L_k(\hat{\eta}_{i-1}) = 0$, $k \ge i$. So $i < j \Rightarrow \tilde{T}_j^i = 0$ while

$$\tilde{T}_i^i = -r^{-1}|\tilde{T}^i|^{-1}L_i(\hat{\eta}_i). \tag{3.27}$$

Now write

$$\begin{aligned} L_i(\hat{\eta}_i) &= L_{ib}(\hat{\eta}_{i-1})\left(\hat{\eta}_i^b - \hat{\eta}_{i-1}^b\right) \\ &\quad + \sum_{k=1}^{m-1}\frac{L_{ib_1\cdots b_k}(\hat{\eta}_{i-1})}{k!}\prod_{j=1}^{k}\left(\hat{\eta}_i^{j} - \hat{\eta}_{i-1}^{j}\right) \\ &\quad + O\left(|\hat{\eta}_{i-1}^i - \hat{\eta}^i|^{m+2}\right) \\ &= \sum_{k=1}^{m-1} P_k(rh)d^k + O(d^{m+2}) \end{aligned} \tag{3.28}$$

by (3.13), where $d = \hat{\eta}_{i-1}^i - \eta^i$ and P_k are polynomials in rh such that

$P_1(0) = 1$. Now apply (3.21) and (3.28) to (3.27) and then substitute (3.22) for d and (ii) follows.

(iii) Follows from Lemma A1 of the Appendix. □

PROOF OF THEOREM 1. By Lemma 3 formula (3.8) is valid for $(\mathbf{x}, t) \in \tilde{S}$. Moreover, from Lemma 2,

$$h^i(t) = t^i + r^{-1}P^i(rt) + O(r^{m+1}), \tag{3.29}$$

where P^i is a polynomial of degree $m + 1$ in rt with no constant or linear term and coefficients which are polynomials in $L_{b_1 \cdots b_k}$, $k \le m + 2$. From (3.23) and (3.29)

$$\begin{aligned} \det\left\|h^i_j(t)\right\| &= \det\left\|\tilde{T}^i_j(\hat{\eta} + rh(t))\right\|^{-1} = \prod_{i=1}^{p} \tilde{T}^i_i(\hat{\eta} + rh(t))^{-1} \\ &= \prod_{i=1}^{p} \left(1 + P^i(rh(t))\right)^{-1} + O(r^{m+1}) \\ &= 1 + V(rt) + O(r^{m-1}), \end{aligned} \tag{3.30}$$

where V is a polynomial of degree $m + 1$ in rt with no constant term and coefficients which are polynomials in $L_{b_1 \cdots b_k}$, $k \le m + 2$.

Moreover, from (3.29) and B_m(i),

$$\begin{aligned} \pi\left(\hat{\theta} + r\hat{D}^{-1}h(t)\right) = \pi(\hat{\theta})\Bigg(1 + \frac{\pi_b(\hat{\theta})}{\pi(\hat{\theta})}U^b(rt) + \cdots \\ + \frac{\pi_{b_1 \cdots b_{m-2}}}{\pi(\hat{\theta})}(\hat{\theta})U^{b_1}(rt) \cdots U^{b_{m+2}}(rt)\Bigg) \\ + O\left(r^{m+1}\pi(\theta)\right), \end{aligned} \tag{3.31}$$

where the U^b are polynomials of degree $\le m + 1$ with no constant term. Substituting back (3.30) and (3.31) in (3.8) provides an approximation to the numerator in (3.8) and integrating this we get an approximation to the denominator in (3.8). Together these approximations ensure that

$$E_P\int \left|\pi_T(t|\mathbf{X}) - \phi(t)(1 + Q^*_m(rt, x, \pi))1\left[(t, \mathbf{X}) \in \tilde{S}\right]\right| dt = O(r^{m+1})$$

for a suitable Q^*_m. We get Q_m by dropping all terms of degree $m + 1$ in Q^*_m. The coefficients are evidently polynomials in $L_{b_1 \cdots b_k}(\hat{\eta})$ and $\pi_{b_1 \cdots b_k}/\pi(\hat{\theta})$, $1 \le k \le m + 1$. But the former are polynomials in the elements of $\hat{D}^{-1}$ which are rational functions of $L_{ij}(\hat{\theta})$. Now,

$$\begin{aligned} E_P\int \phi(t)\left[Q_m(rt, \mathbf{x}, \pi) - Q^*_m(rt, x, \pi)\right]1\left[(t, \mathbf{X}) \in \tilde{S}\right] dt \\ = O(r^{m-1}) \end{aligned} \tag{3.32}$$

since for $\mathbf{x} \in S$ all coefficients in both functions are bounded. Further,

$$E_P\int \pi_T(t|\mathbf{X})1\left((t, \mathbf{X}) \notin \tilde{S}\right) dt = P\left[(T, \mathbf{X}) \notin \tilde{S}\right] = O(r^{m+1}) \tag{3.33}$$

by B_m. Finally,

$$E_P\int \phi(t)Q_m(rt, x, \pi)1(\mathbf{X} \in S, |t| \le M^* r^{m-1} \text{ or } |t| \ge r^{-\delta})\,dt = O(r^{m+1})$$

and the theorem follows. □

PROOF OF THEOREM 2 AND COROLLARY 1. Evidently since D and $\tilde{D}$ are simple transforms of T, we need merely check that the approximation to the density of D ($\tilde{D}$, respectively) obtained by applying the usual transformation formula to $\pi_m(\cdot, \mathbf{X})$ agrees with $\prod_{k=1}^{p} c_1(u^k)$ with error $O(r^{m+1})$ for $m = 1, 3$, respectively. This follows readily from Lemmas A2 and A3 in the Appendix if we identify π_m with $g(t)$ for $m = 2, 3$ and note that $R_{jj} = O(n^{-1})$. Relation (2.6) follows from Lemmas A2 and A3. Corollary 1(a) follows immediately from (2.5), while 1(b) follows from (2.6) and Lemma A4. □

PROOF OF THEOREM 3. Evidently $F_m \Rightarrow B_m$ for π satisfying (vii). It is shown in Ghosh, Sinha and Joshi (1982) and Bickel, Götze and van Zwet (1985) that the set of all such π is dense in the set of all priors under weak convergence. Now (2.9) implies that for any π concentrating on a compact, the characteristic function of T satisfies the approximation

$$\begin{aligned}
\int e^{i\nu_j t^j} p_T(t)\,dt &= \int\int e^{i\nu_j t^j} p_T(t|\theta)\pi(\theta)\,d\theta\,dt \\
&= e^{i\nu_j t^j}\phi(t)\left(1 + \sum_{k=1}^{m} r^k \int R_k(t, \theta)\pi(\theta)\,d\theta\right) dt + O(r^{m+1}) \\
&= \exp\left\{-\tfrac{1}{2}\sum_{j=1}^{p} (\nu^j)^2\right\}\left[1 + \sum_{k=1}^{m} r^k \int P_k(\nu, \theta)\pi(\theta)\,d\theta\right] \\
&\quad + O(r^{m+1}),
\end{aligned} \tag{3.34}$$

where $\exp\{-\frac{1}{2}\sum_{j=1}^{p}(\nu^j)^2\}P_k(\nu, \theta)$ is the Fourier transform of $\phi(t)R_k(t, \theta)$, so that the P_k's are also polynomials in ν. On the other hand, Theorem 1 yields

$$\begin{aligned}
&\int \exp\left\{\sum_{j=1}^{p} (\nu^j)^2\right\} p_T(t)\,dt \\
&\quad = E_P\left[\int \exp\left\{\sum_{j=1}^{p} (\nu^j)^2\right\}\pi_m(t, \mathbf{X})1(\mathbf{X} \in S)\,dt\right] + O(r^{m+1}) \\
&\quad = \exp\left\{-\tfrac{1}{2}\sum_{j=1}^{p} (\nu^j)^2\right\} \\
&\qquad \times\left(1 + \sum_{k=1}^{m} r^k t^{b_1} \cdots t^{b_{r'}} E Q_{mb_1 \cdots b_k}(\mathbf{X}, \pi)1(\mathbf{X} \in S)\right) + O(r^{m+1}).
\end{aligned} \tag{3.35}$$

Therefore, multiplying by $\exp\{\frac{1}{2}\Sigma_{j-1}^{p}(\nu^j)^2\}$ we get

$$1 + \sum_{k-1}^{m} r^k \int P_k(\nu, \theta)\pi(\theta)\, d\theta$$

$$= 1 + \sum_{k-1}^{m} r^k c_{b_1 \cdots b_k}(\pi)\nu^{b_1} \cdots \nu^{b_k} + O(r^{m+1}), \tag{3.36}$$

where O is now uniform for $|\nu| \leq M$ by the hypothesis of Theorem 3.

Define, as usual,

$$\Delta_{b_1 \cdots b_p} f(t^1, \ldots, t^p) = \left(\Delta_1^{b_1} \cdots \Delta_p^{b_p}\right) f(t^1, \ldots, t^p),$$

where the $b_j = 0, \ldots, p$, $\Sigma_{j-1}^{p} b_j = l$ and

$$\Delta_k f = f(t^1, \ldots, t^{k-1}, t^k + \varepsilon, t^{k+1}, \ldots, t^p) - f(t^1, \ldots, t^p)$$

and Δ_k^0 represents an operator product. Apply $\Delta_{b_1 \cdots b_p}$ to both sides of (3.36) considered as functions of ν. If $l > m$ we obtain

$$\sum_{j-1}^{m} r^j \varepsilon^{-l} \int \Delta_{b_1 \cdots b_p} P_j(\varepsilon, \theta)\pi(\theta)\, d\theta = O(r^{m-1}\varepsilon^{-l}). \tag{3.37}$$

Let $\varepsilon \downarrow 0$ more slowly than $r^{1/l}$. Then (3.37) yields

$$\int \frac{\partial^p P_k}{\partial^{b_1} u_1 \cdots \partial^{b_p} u_p}(\nu, \theta)\pi(\theta)\, d\theta = 0 \quad \text{for all } \nu, \text{ for all } k \leq m.$$

But by assumption the integrand is continuous in θ. Since π ranges over a dense set we conclude that the integrand vanishes identically in θ. So P_k is a polynomial of degree less than or equal to k and hence so is R_k. □

Theorem 4 and Corollary 3 follow from Theorem 3 in the same fashion as Theorem 2 and Corollary 1 follow from Theorem 1.

Acknowledgments. We thank Ole Barndorff-Nielsen, Ib Skovgaard and Peter McCullagh for some crucial references.

APPENDIX

LEMMA A1. *Suppose f: $C^0 \to R^p$ where C^0 is an open convex set in R^p. Suppose f is differentiable with Hessian $\dot{f}$ and*

$$|\dot{f} - J| < 1, \tag{A1}$$

where J is the identity and $|M|$ is the operator norm on matrices. Then $\dot{f}$ is nonsingular and f is 1–1.

PROOF. By (A1), $\dot{f}$ is nonsingular:

$$\dot{f}^{-1} = J - (\dot{f} - J) + (\dot{f} - J)^2 \cdots .$$

If $f(a) = f(b)$, then

$$0 = \int_0^1 \dot{f}(a + \lambda(b - a))\, d\lambda(b - a)$$

or

$$(b - a) = -\int_0^1 \big(\dot{f}(a + \lambda(b - a)) - J\big)\, d\lambda(b - a).$$

Then, by (A1),

$$|b - a| \le \max_\lambda |\dot{f}(a + \lambda(b - a)) - J||b - a| < |b - a|$$

unless $b = a$.

LEMMA A2. *Let*

$$g(t) = \phi(t)\big(1 + R_i t^i + R_{ij}(t^i t^j - \delta^{ij}) + R_{ijk} t^i t^j t^k\big)$$

be the density of a finite measure μ on R^p where δ^{ij} is Kronecker delta and let

$$g_0(t) = \phi(t)\Big(1 + R_{jj}\big((t^j)^2 - 1\big)\Big)$$

similarly correspond to μ_0. Let $h(t) = (t^i)^2$. Then

$$\mu h^{-1} = \mu_0 h^{-1}.$$

PROOF. The densities of μh^{-1} and $\mu_0 h^{-1}$ at $(|u^1|, \ldots, |u^p|)$ differ by the term

$$2^p E\Big(R_i \varepsilon_i |u^i|^{1/2} + R^*_{ij} \varepsilon_i \varepsilon_j |u^i|^{1/2} |u^j|^{1/2} + \varepsilon_i \varepsilon_j \varepsilon_k R_{ijk} |u^i|^{1/2} |u^0|^{1/2} |u^k|^{1/2}\Big) = 0,$$

where $R^*_{ij} = R_{ij}(1 - \delta_{ij})$ and ε_i are independent ± 1 with probability $\frac{1}{2}$. □

LEMMA A3. *Suppose $\Sigma_j |R_{jj}| = o(1)$. Then*

$$\int g_0(t) - \prod_{j=1}^p (1 + 2R_{jj})^{-1/2} \phi_1\big(t^j (1 + 2R_{jj})^{-1/2}\big)\, dt = O\Big(\sum_j R_{jj}^2\Big),$$

where ϕ_1 is the standard normal density.

PROOF. Taylor expand. □

LEMMA A4. *Suppose $\Sigma_1^k |c_{jj} - 1| = o(1)$ and $Z_j^2 = c_j Y_j^2$ where Y_j's are i.i.d. $N(0, 1)$. Let $U = \Sigma_1^k Y_j^2$ and $V = (\Sigma Z_i^2)(1 + \Sigma(c_j - 1)/k)^{-1}$. Then U and V have the same characteristic function up to $O(\Sigma(c_j - 1)^2)$.*

PROOF. Compute the characteristic function of V, take logarithms and expand. □

REFERENCES

BARNDORFF-NIELSEN, O. E. (1983). On a formula for the conditional distribution of the maximum likelihood estimator. *Biometrika* **70** 343–365.

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1984). Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. *J. Roy. Statist. Soc. Ser. B* **46** 483–495.

BARNDORFF-NIELSEN, O. E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73** 307–322.

BARNDORFF-NIELSEN, O. E. and HALL, P. (1988). On the level error after Bartlett adjustment of the likelihood ratio statistic. *Biometrika* **75** 374–378.

BARTLETT, M. S. (1937). Properties of sufficiency and statistical tests. *Proc. Roy. Soc. Lond. Ser. A* **160** 268–282.

BHATTACHARYA, R. N. and GHOSH, J. K. (1978). Validity of formal Edgeworth expansion. *Ann. Statist.* **6** 434–451.

BICKEL, P. J. (1974). Edgeworth expansions in nonparametric statistics. *Ann. Statist.* **2** 1–20.

BICKEL, P. J., GÖTZE, F. and VAN ZWET, W. R. (1985). A simple analysis of third order efficiency of estimates. In *Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer* (L. Le Cam and R. A. Olshen, eds.) **2** 749–768. Wadsworth, Belmont, Calif.

BILLINGSLEY, P. (1961). *Statistical Inference for Markov Processes*. Univ. Chicago Press, Chicago.

BOX, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika* **36** 317–346.

CHANDRA, T. and GHOSH, J. K. (1979). Valid asymptotic expansion for the likelihood ratio statistic and other perturbed χ^2 variables. *Sankhyā Ser. A* **41** 22–47.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia.

EFRON, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika* **72** 45–58.

GHOSH, J. K., SINHA, B. K. and JOSHI, S. M. (1982). Expansions for posterior probability and integrated Bayes risk. In *Statistical Decision Theory and Related Topics III* (S. S. Gupta and J. O. Berger, eds.) **1** 403–456. Academic, New York.

GÖTZE, F. and HIPP, C. (1978). Asymptotic expansions under moment conditions. *Z. Wahrsch Verw. Gebiete* **42** 67–87.

LAWLEY, D. N. (1956). A general method for approximating to the distribution of the likelihood ratio criteria. *Biometrika* **43** 295–303.

PFANZAGL, J. (1974). Asymptotically optimum estimation and test procedures. In *Proc. Prague Symp. on Asymptotic Statistics* (J. Hájek, ed.) **1** 201–272. Charles Univ., Prague.

STEIN, C. (1985). On the coverage probability of confidence sets based on a prior distribution. *Sequential Meth. Statist.: Banach Center Publication* **16** 485–514.

WALD, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* **54** 426–482.

WELCH, B. N. and PEERS, B. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. Ser. B* **35** 318–329.

WILKS, S. S. (1938). The large sample distribution of the likelihood ratio statistic for testing composite hypotheses. *Ann. Math. Statist.* **9** 60–62.

DEPARTMENT OF STATISTICS
STATISTICAL LABORATORY
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720

INDIAN STATISTICAL INSTITUTE
203 BARRACKPORE TRUNK RD.
CALCUTTA 700 0–35
INDIA