The Annals of Statistics
1990, Vol. 18, No. 2, 779–789

ON THE DENSITY OF MINIMUM CONTRAST ESTIMATORS

BY IB M. SKOVGAARD

Royal Veterinary and Agricultural University

Conditions for the existence of the density of a minimum contrast estimator in a parametric statistical family are given together with a formula for this density. The formula is exact if multiple local minima cannot occur; otherwise the formula is an exact expression for the point process of local minima of the contrast function. Although it is not in general feasible to compute the expression for the density, the formula can be used as a basis for further expansion of the large deviation type. When the estimate is sufficient, either in the original model or after conditioning on an approximate or exact ancillary, the formula simplifies drastically. In particular, it is shown how Barndorff-Nielsen's formula for the density of the maximum likelihood estimator given an ancillary statistic is derived from the formula given here. In this way the nature of Barndorff-Nielsen's formula as an asymptotic approximation and its appearance as an exact formula for certain cases are demonstrated.

1. Introduction. The theorem given in this paper provides conditions for the existence of the density of the maximum likelihood estimator of a (vector) parameter, together with a formula for this density. The formula is an exact expression of the intensity of the point process of local maxima of the likelihood function, which is the same as the density of the maximum likelihood estimator if multiple local maxima cannot occur; otherwise it exceeds this density by an amount stemming from cases when a local maximum occurs that is not global. Typically, in the case of n independent replications, this difference decreases at an exponential rate in n. The formula was given as Lemma 4.2 in Skovgaard (1985b) under much stronger conditions, within the framework of exponential families. There it was used as an intermediate step toward a large deviation expansion of the density of the maximum likelihood estimator, which turned out to be identical to the one derived by Field (1982) by quite different methods. Field, working with the more general class of M-estimators, assumed the existence of the density of the estimator and proceeded to derive the large deviation expansion by a combination of the methods of Edgeworth expansions from Bhattacharya and Ghosh (1978) and a saddlepoint expansion for the score statistic evaluated at the point at which the density of the estimator was to be calculated. Pazman (1984) derived the same formula for the case of nonlinear regression along a third line of reasoning [cf. Hougaard (1985)]. In the present paper we shall consider the kind of continuity conditions required for the existence of the density of the estimator, a problem that was not addressed for more general families of distributions in any of the

Received April 1988; revised February 1989.
AMS 1980 *subject classifications*. Primary 62F12; secondary 62E15.
Key words and phrases. Barndorff-Nielsen's formula, conditional inference, large deviation expansion, minimum contrast estimator, maximum likelihood estimator, saddlepoint approximation.

papers mentioned above. Another purpose of the present paper is to consider the exact formula in itself, because it may be used to derive more refined approximations for special classes of models, in particular in connection with conditional inference. Thus, as a main example, Barndorff-Nielsen's formula (3.1) [cf. Barndorff-Nielsen (1980, 1983)] is shown to be an easy consequence and its exactness properties may be more transparent in view of the formula given here.

In Section 2 the theorem on the existence and the formula for the density for the estimator are given. The result is stated in terms of minimum contrast estimators, because the method of proof does not rely on any particular properties of the likelihood function. The proof of the theorem is given at the end of that section. Some readers may want to skip this and go directly to the next section.

In Section 3 we discuss the implications of the formula, mainly for conditional inference, in which case it simplifies drastically. In particular, Barndorff-Nielsen's formula is discussed in this section.

2. A formula for the density. Let $\{P_\beta; \beta \in B \subseteq \mathbb{R}^p\}$ be a family of probability measures on some measurable space E, where the parameter space B is some subset of $\mathbb{R}^p$. Consider an estimator $\hat{\beta}$ of the parameter β that minimizes the contrast function

$$-\gamma(y;\beta), \qquad y \in E, \beta \in B,$$

as a function of β when $y \in E$ is the observed data point. For analogy with the special case when γ is the log-likelihood function we prefer to consider the maximization of the negative of the contrast function. The function γ: $E \times \mathbb{R}^p \to \mathbb{R}$ is required to be measurable with respect to the product of P_β and the Borel measure on $\mathbb{R}^p$ for any β. For any $\varepsilon > 0$ and $v \in \mathbb{R}^p$, define

$$U_\varepsilon(v) = \{x \in \mathbb{R}^p; \|x - v\| < \varepsilon\} \tag{2.1}$$

as the ε-neighbourhood of v in terms of the usual Euclidean norm and let $A_p = \text{vol}(U_1(0))$ denote the volume of the unit ball. Given a fixed parameter value $\beta_0 \in B$, we shall consider the existence of the β_0-density of $\hat{\beta}$ at another fixed point $b \in \text{int}(B)$. Thus all probabilities, expectations, densities, etc., in the sequel are, unless otherwise stated, computed with respect to P_{β_0}. Define

$$g(b;\beta_0) = \lim_{\varepsilon \to 0} (\varepsilon^p A_p)^{-1} P\{\gamma(y;\beta) \text{ has a local maximum in } U_\varepsilon(b)\}, \tag{2.2}$$

which is the intensity of the process of local maxima of the function $\gamma(y;\beta)$ at the point $\beta = b$. It may also be interpreted as the density of a local minimum contrast estimator or, if multiple maxima can be ruled out, the density of the global minimum contrast estimator. Let $D_1(\beta)$ denote the column vector of first partial derivatives of $\gamma(y;\beta)$ with respect to the coordinates of β and $D_2(\beta)$ denote the matrix of second derivatives. The abbreviations D_1 and D_2

are used for the values at the point $\beta = b$. We shall need the following conditions.

(C1) With probability 1, $\gamma(y;\beta)$ is three times differentiable at $\beta = b$.

(C2) Conditional densities $f(d_1|d_2;\beta_0)$ of D_1 given D_2 with respect to the Lebesgue measure exist and satisfy

$$f(d_1|d_2;\beta_0) \le F < \infty,$$

for almost all d_1 and d_2, where F is some constant.

(C3) Let

$$M(\varepsilon) = \sup\left\{\left|(d/dh)^3\gamma(y;\beta + hv)\right|; \beta \in U_\varepsilon(b), v \in \mathbb{R}^p, \|v\| = 1\right\},$$

where the derivative is evaluated at $h = 0$, be the maximal norm of the third differential of $\gamma(y;\beta)$ in the ε-neighbourhood of b. Then two constants, $\eta > 0$ and $\zeta > 0$ exist, such that $0 < \zeta + (p-1)\eta < 1$ and that for any $v \in \mathbb{R}^p$,

$$E\left\{\left|v^{\mathsf{T}}D_2 v\right|^{p/\eta}\right\} < \infty,$$

where v^{T} denotes the transpose of v and

$$E\left\{[M(\varepsilon)]^{p/\zeta}\right\} < \infty.$$

THEOREM. *Assume that the conditions* C1, C2 *and* C3 *hold. Then* $g(b;\beta_0)$ *exists and*

$$\begin{aligned} g(b;\beta_0) &= E\left\{f(0|D_2;\beta_0)|D_2|I_{\text{neg}}(D_2)\right\} \\ &= f_1(0;\beta_0)E\left\{|D_2|I_{\text{neg}}(D_2)\,\middle|\,D_1 = 0\right\}, \end{aligned} \tag{2.3}$$

where $|D_2|$ *denotes the absolute value of the determinant of* D_2, $I_{\text{neg}}(D_2)$ *is the indicator function that equals one if* D_2 *is negative definite and zero otherwise, and* f_1 *is the marginal density of* D_1 *with respect to the Lebesgue measure.*

The conditions as stated in the theorem are not optimal. Some minor technical improvements have been sacrificed to avoid obscuring the theorem too much. For example, $M(\varepsilon)$ needs to be well defined only with probability $1 - o(\varepsilon^p)$ as $\varepsilon \to 0$ and the conditional density in C2 needs only to be uniformly bounded for d_1 in some neighbourhood of zero. In fact, the existence of this density with respect to the Lebesgue measure is required only in such a neighbourhood. It is easy to see from the proof that these relaxed conditions are sufficient.

PROOF OF THE THEOREM. Define the function

$$\tilde{\gamma}(y;\beta) = \gamma(y;b) + D_1^{\mathsf{T}}(\beta - b) + \tfrac{1}{2}(\beta - b)^{\mathsf{T}}D_2(\beta - b) \tag{2.4}$$

and the events

$$L(\varepsilon) = \{y \in E; \gamma(y;\beta) \text{ has a local maximum in } U_\varepsilon(b)\}, \tag{2.5}$$

$$\tilde{L}(\varepsilon) = \{y \in E; \tilde{\gamma}(y;\beta) \text{ has a local maximum in } U_\varepsilon(b)\}. \tag{2.6}$$

The quantity we seek is given in (2.2), but it will appear from the proof that it can also be calculated as the limit

$$\tilde{g}(b;\beta_0) = \lim_{\varepsilon\to 0} (\varepsilon^p A_p)^{-1} P\{\tilde{L}(\varepsilon)\}, \tag{2.7}$$

which is identical to (2.2) except that $L(\varepsilon)$ is replaced by $\tilde{L}(\varepsilon)$. It follows from the condition C3 by use of Chebyshev's inequality that the event

$$S(\varepsilon) = \{y \in E; \|D_2\| \le \varepsilon^{-\tilde{\eta}} \text{ and } M(\varepsilon) \le \varepsilon^{-\tilde{\zeta}}\} \tag{2.8}$$

has probability $1 - o(\varepsilon^p)$ as $\varepsilon \to 0$, where $\tilde{\eta} > \eta$ and $\tilde{\zeta} > \zeta$ are two fixed constants chosen to satisfy $\tilde{\zeta} + (p-1)\tilde{\eta} < 1$ and $\|D_2\|$ is the largest absolute eigenvalue of D_2. Within the event $S(\alpha)$ we have the estimate

$$|\gamma(y;\beta) - \tilde{\gamma}(y;\beta)| \le \tfrac{1}{6}\alpha^3 M(\alpha) \le \tfrac{1}{6}\alpha^3\alpha^{-\tilde{\zeta}}, \tag{2.9}$$

for any $\beta \in U_\alpha(b)$. Therefore, if $y \in \tilde{L}(\varepsilon) \cap S((1+\delta)\varepsilon)$ such that the function $\tilde{\gamma}(y;\beta)$ has a local maximum at β_1, say, in $U_\varepsilon(b)$, then for some sufficiently small $\delta > 0$ and $v \in \mathbb{R}^p$ with $\|v\| = 1$ we have

$$\begin{aligned} &\gamma(y;\beta_1) - \gamma(y;\beta_1 + \delta\varepsilon v) = -\tfrac{1}{2}\delta^2\varepsilon^2 v^{\mathsf{T}} D_2 v + (\gamma(y;\beta_1) - \tilde{\gamma}(y;\beta_1)) \\ &-(\gamma(y;\beta_1 + \delta\varepsilon v) - \tilde{\gamma}(y;\beta_1 + \delta\varepsilon v)) = -\tfrac{1}{2}\delta^2\varepsilon^2 v^{\mathsf{T}} D_2 v \pm \tfrac{1}{3}(1+\delta)^3\varepsilon^{3-\tilde{\zeta}}, \end{aligned} \tag{2.10}$$

where we have used (2.9) with $\alpha = \varepsilon(1+\delta)$. Notice that on the set $\tilde{L}(\varepsilon)$ the matrix D_2 must be negative semidefinite and hence the first term on the right in (2.10) is nonnegative. We want to infer that (2.10) is, in fact, positive for any v when ε is sufficiently small and hence that $\gamma(y;\beta)$ is larger at β_1 than anywhere on the boundary of $U_{\delta\varepsilon}(\beta_1)$, implying that $L(\varepsilon(1+\delta))$ has occurred. To do this we must be able to exclude the event

$$R_c(\varepsilon) = \bigcup_{\|v\|=1} \{y \in E; |v^{\mathsf{T}} D_2 v| \le c\varepsilon^{1-\tilde{\zeta}}\} \tag{2.11}$$

for $c = c(\delta) = \frac{2}{3}\delta^{-2}(1+\delta)^3$. On this set the smallest absolute eigenvalue of D_2 is bounded by $c\varepsilon^{1-\tilde{\zeta}}$, while the remaining eigenvalues are bounded on the set $S((1+\delta)\varepsilon)$ by $\varepsilon^{-\tilde{\zeta}}$. Hence, if $y \in \tilde{L}(\varepsilon) \cap S((1+\delta)\varepsilon) \cap R_c(\varepsilon)$, then

$$|D_2| \le c\varepsilon^{1-\tilde{\zeta}}(\varepsilon^{-\tilde{\zeta}})^{p-1} = c\varepsilon^{1-\tilde{\zeta}-(p-1)\tilde{\eta}} = o(1) \tag{2.12}$$

as $\varepsilon \to 0$ for any $c > 0$. If $\tilde{\gamma}(y;\beta)$ has a local maximum in $U_\varepsilon(b)$ we must have

$$D_1 \in -D_2(U_\varepsilon(0)), \tag{2.13}$$

which is the set of vectors $-D_2 v$ with $v \in U_\varepsilon(0)$. This set has the volume

$$\text{vol}\{-D_2(U_\varepsilon(0))\} = |D_2|\varepsilon^p A_p. \tag{2.14}$$

By use of the assumption that the conditional density of D_1 given D_2 is bounded, it follows immediately from (2.12), (2.13) and (2.14) that

$$P\big(\tilde{L}(\varepsilon) \cap S((1+\delta)\varepsilon) \cap R_c(\varepsilon)\big) = o(\varepsilon^p) \tag{2.15}$$

as $\varepsilon \to 0$ for any $c > 0$. Together with the fact that $S((1+\delta)\varepsilon)$ has probability $1 - o(\varepsilon^p)$ as $\varepsilon \to 0$, this implies that

$$\begin{aligned} \tilde{g}(b;\beta_0) &= \lim_{\varepsilon\to 0} \big(\varepsilon^p A_p\big)^{-1} P\{\tilde{L}(\varepsilon)\} \\ &= \lim_{\varepsilon\to 0} \big(\varepsilon^p A_p\big)^{-1} P\big\{\big(\tilde{L}(\varepsilon) \cap S((1+\delta)\varepsilon)\big) \setminus R_c(\varepsilon)\big\}. \end{aligned} \tag{2.16}$$

It now follows from (2.10) that

$$\big(\tilde{L}(\varepsilon) \cap S((1+\delta)\varepsilon)\big) \setminus R_{c(\delta)}(\varepsilon) \subseteq L(\varepsilon(1+\delta))$$

and hence that

$$\begin{aligned} \tilde{g}(b;\beta_0) &\le \lim_{\varepsilon\to 0} \big(\varepsilon^p A_p\big)^{-1} L(\varepsilon(1+\delta)) \\ &= (1+\delta)^p \lim_{\varepsilon\to 0} \big(\varepsilon^p (1+\delta)^p A_p\big)^{-1} L(\varepsilon(1+\delta)) \\ &= (1+\delta)^p g(b;\beta_0). \end{aligned}$$

Since this holds for any (sufficiently small) $\delta > 0$ we have proved that

$$\tilde{g}(b;\beta_0) \le g(b;\beta_0). \tag{2.17}$$

The other inequality is proved along the same lines although some of the arguments are slightly different. Thus, assume that $y \in L(\varepsilon)$ and let $\beta_1 \in U_\varepsilon(b)$ be a local maximum point of $\gamma(y;\beta)$. Then the derivative of γ with respect to β at β_1 is zero, i.e.,

$$\begin{aligned} 0 &= \frac{d}{d\beta}\gamma(y;\beta_1) \\ &= D_1 + D_2(\beta_1 - b) \pm \tfrac{1}{2}\varepsilon^2 M(\varepsilon) \\ &= D_1 + D_2(\beta_1 - b) \pm \tfrac{1}{2}\varepsilon^{2-\tilde{\zeta}}, \end{aligned}$$

if $y \in S(\varepsilon)$, where the notation $\pm$ is used for any vector with a length limited by the quantity indicated. Hence

$$D_1 \in -D_2(U_\varepsilon(0)) \pm \tfrac{1}{2}\varepsilon^{2-\tilde{\zeta}}.$$

As in the first half of the proof we want to exclude the possibility that D_2 has numerically small eigenvalues, i.e., the set $R_c(\varepsilon)$ from (2.11) for some appropriate c. As above an occurrence of the event $R_c(\varepsilon)$ implies that

$$\begin{aligned} \operatorname{vol}\big\{-D_2(U_\varepsilon(0)) \pm \tfrac{1}{2}\varepsilon^{2-\tilde{\zeta}}\big\} &\le \big(c\varepsilon^{2-\tilde{\zeta}} + \tfrac{1}{2}\varepsilon^{2-\tilde{\zeta}}\big)\big(\varepsilon^{1-\tilde{\eta}} + \tfrac{1}{2}\varepsilon^{2-\tilde{\zeta}}\big)^{p-1} \\ &\le c_1 \varepsilon^{p+1-\tilde{\zeta}-(p-1)\tilde{\eta}} = o(\varepsilon^p) \end{aligned} \tag{2.18}$$

as $\varepsilon \to 0$, where $c_1 > 0$ is some constant. As in calculation (2.16), this shows

that the set $R_c(\varepsilon)$ may be ignored for any $c > 0$. For any $v \in \mathbb{R}^p$ with $\|v\| = 1$ and any $y \in S(\varepsilon)$, the inequality

$$\left|v^{\mathrm{T}}D_2(\beta_1)v - v^{\mathrm{T}}D_2 v\right| \leq \varepsilon M(\varepsilon) \leq \varepsilon^{1-\zeta} \tag{2.19}$$

shows that also the set on which $D_2(\beta_1)$ has any eigenvalue less than or equal to $c\varepsilon^{1-\zeta}$ may be ignored because it has probability $1 - o(\varepsilon^p)$. We still have the bound (2.9) for any $y \in S(\alpha)$ and therefore for any $y \in (L(\varepsilon) \cap S((1+\delta)\varepsilon)) \setminus R_c(\varepsilon)$ and any v with $\|v\| = 1$:

$$\begin{aligned}
&\tilde{\gamma}(y;\beta_1) - \tilde{\gamma}(y;\beta_1 + \delta\varepsilon v) \\
&\quad = \gamma(y;\beta_1) - \gamma(y;\beta_1 + \delta\varepsilon v) \pm \tfrac{1}{3}(1+\delta)^3\varepsilon^{3-\zeta} \\
&\quad = -\tfrac{1}{2}\delta^2\varepsilon^2 v^{\mathrm{T}}D_2(\beta_1)v \pm \left\{\tfrac{1}{3}(1+\delta)^3\varepsilon^{3-\zeta} + \tfrac{1}{6}(\delta\varepsilon)^3\varepsilon^{-\zeta}\right\} \\
&\quad > 0
\end{aligned} \tag{2.20}$$

if the constant c is chosen appropriately. Thus, it follows that $\tilde{\gamma}(y;\beta)$ has a local maximum in the set $U_{\delta\varepsilon}(\beta_1) \subseteq U_{\varepsilon(1+\delta)}(b)$ and the inequality opposite to (2.17) is derived by the argument analogous to the one leading to (2.17). We conclude that

$$g(b;\beta_0) = \tilde{g}(b;\beta_0). \tag{2.21}$$

On the set of negative definite D_2's the local maximum of $\tilde{\gamma}(y;\beta)$ is located at $b - D_2^{-1}D_1$, which has the conditional density

$$|D_2| f(d_1 | D_2; \beta_0)$$

at the point $b - D_2^{-1}d_1$, given D_2. From the arguments above it is seen that the set of singular D_2's, as a subset of $R_c(\varepsilon)$, does not contribute to the limit (2.7) and, therefore, the density $\tilde{g}(b;\beta_0)$ of the local maximum of $\tilde{\gamma}(y;\beta)$ at b equals the first expression in (2.3). The second expression in (2.3) is a simple recast of the first. □

3. Application to conditional inference. In this section we demonstrate the use of (2.2) to derive approximate formulas for the density of the estimator. In particular it will appear that Barndorff-Nielsen's formula for the conditional density for the maximum likelihood estimator given an ancillary statistic may be derived from (2.3). In the sequel we abandon any kind of rigour in the sketched proofs. The point is to show the potential of the theorem, not to provide new proofs of known results. In particular the technical problems related to the possibility of multiple local maxima are ignored. Throughout the section we consider only maximum likelihood estimation, i.e., the function $\gamma(y;\beta)$ is the logarithm of the density of y at β. We shall denote this function by $l(\beta;y)$ in the sequel. The likelihood considered is the likelihood from the model, even when we consider conditional inference, in which case an alternative would have been to maximize the conditional likelihood. In particular, it should be noticed that the functions $D_1(\beta)$ and $D_2(\beta)$ are the first two derivatives of the log-likelihood, unaffected by conditioning. However, we may want a formula for the conditional density of the

(unconditional) maximum likelihood estimator given some exact or approximate ancillary statistic; then the theorem of the previous section applies to the conditional distributions, i.e., in the conditions and the resulting formula (2.3), all densities and expectations are conditioned on the ancillary.

Barndorff-Nielsen's formula [cf. Barndorff-Nielsen (1980, 1983)] for the conditional density of $\hat{\beta}$ at the point b given the event $A(y) = a$, where $A = A(y)$ is an exact or approximate ancillary statistic, is

$$(3.1)s \qquad g(b|a;\beta_0) \sim c(a)\,|\,j(b;y)\,|^{1/2} \exp\{l(\beta_0;y) - l(b;y)\},$$

where $c(a)$ is a constant depending only on a, $j(\beta;y)$ equals $-D_2(\beta)$ and $y = y(b,a)$ is any data value for which $A(y) = a$ and $\hat{\beta}(y) = b$. Thus, for the formula to be well defined it is required that the joint statistic $(\hat{\beta}, A)$ be sufficient, in which case the formula does not depend on which of the possible data values y is chosen. The constant $c(a)$ is usually taken either as the general approximation $(2\pi)^{-p/2}$ or as the normalizing constant for which the integral of g with respect to b is 1. Some virtues of the formula are that it is exact for transformation models, it is equivalent to a saddlepoint approximation for full exponential family, it is an accurate approximation for subfamilies of exponential families and the formula applies to a large class of ancillary statistics. Several proofs of its validity as an asymptotic approximation for curved exponential families have been given [cf. Barndorff-Nielsen (1980), McCullagh (1987, Section 8.6) and Fraser (1988)], so the reader may well question the desirability of yet another. However, the main point of the following derivation, apart from its simplicity, is that it sheds some light over some features of the formula (3.1), including cases of exactness, its nature as an asymptotic approximation, its relation to the saddlepoint approximation and the appearance of the observed rather than the expected Fisher information in the formula. In the original proof of the general asymptotic nature of the approximation in Barndorff-Nielsen (1980), the observed information remained to some extent an optional choice compared to the expected information. This choice was, however, fully justified by a number of cases for which the formula turned out to be exact.

The discussion below is related, in particular, to the one in Fraser (1988). A difference is that Fraser, as McCullagh (1987), works with the score function at the point β_0; another is that the formula (2.3) applies also to cases of nonsufficiency for which other saddlepoint-type expansions may be derived from it [cf. Skovgaard (1985b)]; a third difference is that we focus our attention on the nature of (3.1) as an asymptotic expansion of the large deviation type.

To see how (3.1) follows from the theorem in Section 2, let us rewrite the last expression in (2.3) as a product of three factors,

$$(3.2) \qquad \begin{aligned} g(b|a;\beta) = {} & \{f(0|a;b)\}\{f(0|a;\beta_0)/f(0|a;b)\} \\ & \times \{E_{\beta_0}(|D_2|I_{\text{neg}}(D_2)|D_1 = 0, a)\}, \end{aligned}$$

where the conditioning on $A = a$ is included as opposed to (2.3). Now, as is always the case when (3.1) is applicable, we require $(\hat{\beta}, A)$ to be sufficient.

Then, for a fixed b, (D_1, A) will also be sufficient because of the one-to-one correspondence, at least locally, between these two pairs of statistics. From Neyman's factorization criterion it follows that D_2 is a function of (D_1, A) and hence that the conditional expectation in the third factor in (3.2) trivially equals $|D_2| = |j(b;y)|$. Thus, we may rewrite (3.2) as

$$g(b|a;\beta) = \{f(0|a;b)\}\{f(0|a;\beta_0)/f(0|a;b)\}|\, j(b;y)\,|. \tag{3.3}$$

We now discuss the reduction of (3.3) to the approximation in (3.1). We consider the case when A is exactly ancillary and then the general case of an approximate ancillary. We also discuss the two cases of special interest, namely the full exponential families and the transformation models.

Full exponential families. In the cases of full exponential families the formula (3.1) is known to be equivalent to a saddlepoint expansion for the density of the score statistic. In this case there is no conditioning since the maximum likelihood estimator $\hat{\beta}$ is itself sufficient. Hence, recalling the sufficiency of the score function D_1 at any parameter value for this class of models, the second factor in (3.3) equals $\exp\{l(\beta_0;y) - l(b;y)\}$.

The mean of D_1 is zero and its variance is $-D_2$, which equals the observed as well as the expected information. Therefore the normal approximation to the P_b-density of D_1 at zero is

$$f(0;b) \sim (2\pi)^{-p/2}|\operatorname{var}_b(D_1)|^{-1/2} = (2\pi)^{-p/2}|\, j(b;y)\,|^{-1/2}. \tag{3.4}$$

Thus (3.1) follows from the approximation (3.3). Notice that the only approximation involved is (3.4), which is a normal approximation to a density at its mean. This approximation is identical to a saddlepoint approximation to the density of D_1 at zero and results in a relative error of $O(n^{-1})$ in (3.4) and consequently also in (3.1) for the case of n independent replications.

Exact ancillaries. If $A(y)$ is exactly ancillary, as is the case in transformation models where the maximal invariant statistic is used, the second factor in (3.3) equals $\exp\{l(\beta_0;y) - l(b;y)\}$. The first factor is the density at zero of the score statistic D_1 at the point b in terms of the conditional distribution induced by the parameter value b. The expected value in this distribution is zero, and consequently the normal approximation to the density at this point becomes

$$f(0|a;b) \sim (2\pi)^{-p/2}|\operatorname{var}_b(D_1|a)|^{-1/2}. \tag{3.5}$$

Viewed together with the separation of the first two factors in (3.3), this approximation is entirely identical to a saddlepoint approximation as was the case for the exponential family.

It remains to obtain an approximation to the conditional variance of D_1 given A. As argued above, D_2 is a function of D_1 and A. Consider the case of n independent replications and assume that the ancillary statistic is in one-to-one correspondence with a function of the minimal sufficient statistic and that

this function does not depend on the number of observations. This is the case if A is chosen, e.g., as the affine ancillary in a curved exponential family [cf. Barndorff-Nielsen (1980)]. Then, perhaps after a one-to-one transformation of A, we may write

$$n^{-1}D_2 = h\big(n^{-1}D_1, A\big),$$

where h is a function that is independent of n. Under mild conditions this function may be expanded as

$$\frac{1}{n}D_2 \sim h(0, A) + \frac{1}{n}h'(0, A)D_1 + \frac{1}{2n^2}D_1^{\mathrm{T}}h''(0, A)D_1 + \cdots, \tag{3.6}$$

where, in a formal notation valid for each coordinate of D_2, we have used h' and h'' to denote the first two derivatives of h with respect to its first argument. If we take expectations on the right in (3.6) for fixed A, we see that the expectation of D_2 equals its value $D_2(0, A)$ as a function of $D_1 = 0$ and A, apart from an error term which is $O(n^{-1})$. But the expectation of $-D_2$ in the conditional distribution equals the conditional variance of D_1 if A is exactly ancillary, and therefore the conditional variance of D_1 may be approximated by $-D_2(0, a)$ which is identical to $j(b; y)$. On combination with (3.3) and (3.5), we arrive at (3.1) with $c = (2\pi)^{-p/2}$. Notice that the argument related to (3.6) shows why the observed rather than the expected Fisher information appears in the formula.

Transformation models. For the case of a transformation model the ancillary statistic is the maximal invariant which is exactly ancillary. From the discussion above it follows that the only remaining step to prove that (3.1) is exact is to prove the transformation invariance of the expression

$$|\, j(b; y)|^{1/2} f(0|a; b)$$

for $y = y(b, a)$, i.e., $b = \hat{\beta}(y)$ and $a = A(y)$. Then it follows that this expression is a function of a alone and (3.1) is therefore exact if $c(a)$ is chosen correctly. To avoid the theory of group actions we shall not verify this invariance; the arguments are given in Barndorff-Nielsen (1983). The reader may easily check the result in the case of a location model for independent identically distributed random variables, even if the model is reparametrized from the location parameter to a smooth function of this.

Approximate ancillaries. So far we have neglected the problem that A need not be exactly ancillary and as a consequence the only approximation involved was to the first factor in (3.3), based on the expansion (3.6). This approximation is of the large deviation type, i.e., it involves an error, here $O(n^{-1})$, which is added to a term that is bounded away from zero for b in a compact set. Thus, the resulting error is a relative error of this order of magnitude as $n \to \infty$. In particular, the relative error remains of this order of magnitude if b is kept fixed as $n \to \infty$, which is a sequence of large deviations in the standardized distribution of $\hat{\beta}$. The saddlepoint expansion has the same

characteristics whereas, e.g., Edgeworth expansions provide bounds only on the additive errors to the density. The approximation derived above, for exact ancillaries, is not normalized. If the error involved is differentiable as a function of b, a renormalization obtained by dividing by the integral over some compact set will reduce the order of magnitude of the additive error to $O(n^{-3/2})$.

Compared to the case of an exact ancillary, three more approximations are generally required. The second factor in (3.3) is not exactly equal to the ratio between the marginal likelihoods. Instead it equals the ratio of the conditional likelihoods, and consequently the error introduced by use of the marginal likelihoods equals the ratio of the densities of A at a with respect to the parameters b and β_0. If A is a first order ancillary, e.g., as the Efron and Hinkley (1978) ancillary, this ratio is, by definition, $1 + O(n^{-1/2})$. A renormalization again improves the error of the resulting approximation to $O(n^{-1})$. For a second order ancillary the error of the approximation is $O(n^{-1})$, or $O(n^{-3/2})$ when renormalized. It should be noticed, however, that the approximation only applies to normal deviates of the ancillary A whenever this is only approximately ancillary since we have no control of the ratio between the two densities of A at a outside a set of normal deviations. If extreme accuracy is required, a remedy is to use the conditional likelihoods in (3.1) [cf. Barndorff-Nielsen (1983)].

The other two inaccuracies that appear when A is only approximately ancillary occur in the approximations to the conditional mean and variance of D_1 given $A = a$. If A is a first order asymptotic ancillary, then the variance of the conditional mean of D_1 given A is $O(1)$; see, e.g., Skovgaard (1985a). Hence the conditional mean, having mean zero over the distribution of A, is itself $O(1)$ and since the variance of D_1 is of order n, the error involved in the normal density approximation (3.5), due to the bias, is $O(n^{-1})$.

For the conditional variance of D_1 given A we notice that the argument based on the expansion (3.6) is still valid for a first order approximate ancillary, but the linear term in D_1 in (3.6) now contributes an amount $O(n^{-1/2})$ to the expectation, and an amount of the same order of magnitude is due to the fact that the variance of D_1 given A is no longer exactly equal to the conditional mean of $-D_2$. If the ancillary is of second order, both of these contributions to the error become $O(n^{-1})$.

It should be noticed that even for an asymptotic ancillary statistic that is only first order ancillary, the resulting expansion (3.1) is still of the large deviation type in the sense that it keeps a bounded relative error in a fixed interval around β_0. Therefore the order of magnitude of the additive error can be improved to $O(n^{-1})$ by renormalization. The error is also improved to this order of magnitude, without renormalization, if we restrict attention to normal deviates of $\hat{\beta}$, i.e., to values of b within a neighbourhood of size $O(n^{-1/2})$ around β_0. This is so because any first order ancillary is a local ancillary of second order [cf. Cox (1980) and Skovgaard (1985a)]. In this connection it is of interest to notice that Amari and Kumon (1983) have proved that in a (k, p) exponential family, a second order ancillary of the type in (3.6) exists, but in

general no higher order ancillaries that are independent of the sample size as required in (3.6).

The general conclusion is that the formula (3.1), like the saddlepoint expansion, has the features of a large deviation type expansion, namely by keeping a uniform relative error over a fixed interval of parameter values. If the approximate ancillary is of first order, the error is $O(n^{-1/2})$; if it is of second order the error is $O(n^{-1})$. In any case the additive error is improved by one order of magnitude by renormalization. However, the formula applies only to normal deviates of the ancillary A unless this is exactly ancillary.

Acknowledgments. This manuscript was completed while the author was on leave at the University of British Columbia. Partial support was provided by the Natural Science and Engineering Research Council of Canada and the Danish Natural Science Research Council.

REFERENCES

AMARI, S. and KUMON, M. (1983). Differential geometry of Edgeworth expansions in curved exponential family. *Ann. Inst. Statist. Math.* **35** 1–24.

BARNDORFF-NIELSEN, O. E. (1980). Conditionality resolutions. *Biometrika* **67** 293–310.

BARNDORFF-NIELSEN, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70** 343–356.

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

COX, D. R. (1980). Local ancillarity. *Biometrika* **67** 279–286.

EFRON, B. and HINKLEY, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* **65** 457–487.

FIELD, C. (1982). Small sample asymptotic expansions for multivariate *M*-estimates. *Ann. Statist.* **10** 672–689.

FRASER, D. A. S. (1988). Normed likelihood as saddlepoint approximation. Technical Report No. 1, Univ. Toronto.

HOUGAARD, P. (1985). Saddlepoint approximations for curved exponential families. *Statist. Prob. Lett.* **3** 161–166.

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

PAZMAN, A. (1984). Probability distribution of the multivariate nonlinear least squares estimates. *Kybernetika* **20** 209–230.

SKOVGAARD, I. M. (1985a). A second-order investigation of asymptotic ancillary. *Ann. Statist.* **13** 534–551.

SKOVGAARD, I. M. (1985b). Large deviation approximations for maximum likelihood estimators. *Probab. Math. Statist.* **6** 89–107.

DEPARTMENT OF MATHEMATICS
ROYAL VETERINARY AND AGRICULTURAL UNIVERSITY
THORVALDSENSVEJ 40
DK-1871 FREDERIKSBERG C
DENMARK