Ib M. Skovgaard

Large Deviation Approximations for Maximum Likelihood Estimators

Preprint
July
1981
9

Institute of Mathematical Statistics
University of Copenhagen

Ib M. Skovgaard

LARGE DEVIATION APPROXIMATIONS FOR MAXIMUM LIKELIHOOD ESTIMATORS *

Preprint No. 9

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

July 1981

ABSTRACT. A large deviation expansion of the density of a maximum likelihood estimator is derived in the case of replications from a multivariate curved subfamily of a continuous exponential family. Apart from an exponentially decreasing term, the approximation deviates only by a relative error of order $O(n^{-1})$ from the true density in a fixed neighbourhood of the true parameter value. An example is given which shows an excellent tail approximation even for small n. The results are specialized to the multidimensional non-linear normal regression models, and it is shown, that in these models, the approximation may be improved to deviate only by an exponentially decreasing error term.

Some key words: curved exponential family, density approximation, large deviations, maximum likelihood estimator, non-linear normal regression, saddlepoint approximation.

1. Introduction

We shall derive an approximation to the density of the maximum likelihood estimator (MLE) of a vector parameter β in the case of a smooth subfamily of an exponential family of continuous type. The expansion is a large deviation expansion in the sense, that under simple replications the relative error of the approximation is $O(n^{-1})$ as $n \to \infty$, uniformly in a fixed neighbourhood of the true parameter value. This fact ensures a much better tail approximation to the distribution, than that obtained by an Edgeworth expansion, where the density is only approximated up to a fixed (not relative) error over the whole range. This may be sufficient for large n (or moderate n), but for small (or moderate) n other approximations are needed.

The computational work required to derive the approximation is probably larger than that required to derive the first and may be the second term of the Edgeworth expansion, and also integration of the approximate density will usually have to be done numerically. However, in a very common class of models, namely the non-linear normal regression models, the result may be stated explicitely and is a very simple algebraic expression (see Section 6). Also, in any case, the complexity of the calculations is mainly determined by the dimension of the parameter space and does not increase with increasing sample size. In fact, if n is the number of replications, the approximation to the density g(b) of $\hat{\beta}$ (the MLE), takes the form $\sqrt{n}\, g_1(b) \exp\{-ng_2(b)\}$, where g_1 and g_2 are non-negative functions.

There are several ways of refining the approximation, some of which will be mentioned in the paper, but each, of course, at the prize of an increased amount of numerical work. We shall be concerned mainly with the simplest version.

The main ideas of the approximations are Lemma 4.2, which gives an exact, although not directly computable, expression for the intensity of the process of local maxima of the likelihood function, together with an application of the saddlepoint approximation, see Daniels (1954), to this expression. The paper has been restricted to maximum likelihood estimators within curved exponential families; it will, however, be clear, that the method may be applied to other estimators and other models.

The paper has been restricted to the derivation of the basic expansion with only a few remarks on the (rather obvious) applications. More efficient use of the expansion in the construction of critical regions and confidence regions is probably possible, but a discussion of these problems would be beyond the scope of this paper.

A few notational definitions, needed only for the multivariate algebra in Sections 4,6 and 7, are given in Section 2. In Section 3 we review the basic method, deriving the expansion in the one-dimensional non-linear normal regression model without attention to mathematical rigour. In Section 4, we shall derive the approximation rigorously in the general (multivariate) curved exponential family model, including the basic proofs, but postponing technical proofs to the Appendix. Section 5 contains an example illustrating the behaviour of the approximation for small n. For large n the behaviour can be deduced from the theorems on its asymptotic properties. In Section 6 we obtain the approximation for the important class of multivariate non-linear normal regression models, using the general results of Section 4. Finally, the Appendix contains the more technical proofs, whereas, as mentioned above, the conceptually important proofs are included in Section 4.

2. Notation

Most of the notation will be easily understood or explained, where it occurs. In Sections 4,6 and 7 we shall, however, use a slightly generalized matrix notation. Vectors, matrices and three-dimensional arrays of numbers are all regarded as matrices, e.g. $A = (a_{ijk})$, $i = 1,\dots,n$; $j = 1,\dots,m$; $k = 1,\dots,m$, is an $n \times m \times m$ matrix. If $B = (b_{\alpha\beta})$ is an $m \times n$ matrix; then AB is the matrix product with respect to the last index of A and the first index of B, i.e.

$$(AB)_{ij\beta} = \sum_{k=1}^{m} a_{ijk} b_{k\beta},$$

which is an $n \times m \times n$ matrix; etc. We shall sometimes emphasize the dimensions of a matrix by writing these as subscripts, e.g. $(a_{ijk})_{n \times m \times m}$ for A or $(c_i)_n$ for a vector c in $\mathbb{R}^n$. To make the notation as conventional as possible, we shall still regard vectors as column-vectors (i.e. c is an $n \times 1$ matrix) unless otherwise indicated, and write c' for the transpose of c. Also B' is the transpose of B.

If $f : \mathbb{R}^m \to \mathbb{R}^n$ is a differentiable function, we shall denote its differential by Df, i.e.

$$Df(x) = \left(\frac{df_i}{dx_j}(x)\right)_{n \times m}, \quad x \in \mathbb{R}^m,$$

and similarly $D^2f(x)$ is the $n \times m \times m$ matrix of second partial derivatives at x, etc.

3. The one-dimensional non-linear normal regression model

Let $X_i = \mu_i(\beta) + \varepsilon_i$, $\varepsilon_i \sim N(0,\sigma^2)$, $\beta \in B \subseteq \mathbb{R}$, B open, $\mu_i : B \to \mathbb{R}$ twice continuously differentiable, $i = 1,\dots,k$ and $\varepsilon_1,\dots,\varepsilon_k$ mutually independent. β is the unknown parameter; we shall consider σ^2 to be known, since this makes no difference in estimating β. We assume the existence of the maximum likelihood estimator $\hat{\beta}$ of β, and to avoid technical details we shall also assume, that only one local maximum of the likelihood function can occur. Both of these assumptions will be relaxed in the next section.

Let $\beta_0 \in B$ be a fixed (true) value of the parameter, and let $g_0(b)$ denote the β_0-density of $\hat{\beta}$ at an arbitrary fixed point b. For the various functions of β, we shall use the convention, that if the argument is omitted, b is understood, whereas an index 0 means the function evaluated at β_0. Define

$$D_j(\beta) = \frac{d^j}{d\beta^j} \log f(x;\beta), \quad j = 1,2 \tag{3.1}$$

where $X = (X_1,\dots,X_k)$ and f is the density of X, and let Q_0 be the β_0 - distribution of $(D_1,D_2) = (D_1(b), D_2(b))$. Then a formal computation yields

$$g_0(b) = \lim_{\varepsilon\to 0} (2\varepsilon)^{-1} P_0\{b - \varepsilon \leq \hat{\beta} \leq b + \varepsilon\}$$

$$= \lim_{\varepsilon\to 0} (2\varepsilon)^{-1} P_0\{D_1(\beta) = 0 \text{ and } D_2(\beta) < 0 \text{ for some } \beta \text{ in }]b-\varepsilon, b+\varepsilon[\}$$

$$= \lim_{\varepsilon\to 0} (2\varepsilon)^{-1} P_0\{D_1 + D_2(\beta - b) = 0 \text{ and } D_2 < 0 \text{ for some } \beta \text{ in }]b-\varepsilon, b+\varepsilon[\}$$

$$= \lim_{\varepsilon\to 0} (2\varepsilon)^{-1} \int_{-\infty}^{0} \int_{-\varepsilon|D_2|}^{\varepsilon|D_2|} dQ_0(d_1,d_2)$$

$$= h_0(0) E_0\{|D_2| \cdot I_{\{D_2 < 0\}} \mid D_1 = 0\}$$

$$= h_0(0)\,e_0, \text{ say,} \tag{3.2}$$

where h_0 is the β_0 - density of D_1, and $I_{\{...\}}$ denotes the indicator-function of the set $\{...\}$. Notice, that no approximations are involved in this computation; the general version will be given in Lemma 4.2 and its proof in the Appendix.

To compute (3.2), notice that the distribution of (D_1, D_2) is bivariate normal with parameters

$$E_0\{D_1, D_2\} = (\gamma_1, \gamma_2), \quad V_0\{D_1\} = \sigma_{11}, \quad V_0\{D_2\} = \sigma_{22}, \quad V_0\{D_1, D_2\} = \sigma_{12}$$

given by

$$\gamma_1 = (\mu(\beta_0) - \mu(b))'\left(\frac{d}{d\beta}\mu(b)\right)/\sigma^2$$

$$\gamma_2 = (\mu(\beta_0) - \mu(b))'\left(\frac{d^2}{d\beta^2}\mu(b)\right)/\sigma^2 - I(b)$$

$$\sigma_{11} = I(b) = \left(\frac{d}{d\beta}\mu(b)\right)'\left(\frac{d}{d\beta}\mu(b)\right)/\sigma^2$$

$$\sigma_{22} = \left(\frac{d^2}{d\beta^2}\mu(b)\right)'\left(\frac{d^2}{d\beta^2}\mu(b)\right)/\sigma^2$$

$$\sigma_{12} = \left(\frac{d}{d\beta}\mu(b)\right)'\left(\frac{d^2}{d\beta^2}\mu(b)\right)/\sigma^2 \tag{3.3}$$

where $\mu(\beta) = (\mu_1(\beta), \ldots, \mu_k(\beta))'$, and $I(\beta)$ is the Fisher - information.

A direct computation now gives

$$h_0(0) = (2\pi I(b))^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}\gamma_1^2/I(b)\} \tag{3.4}$$

$$e_0 = \alpha\Phi(\alpha/\tau) + \tau\phi(\alpha/\tau) \tag{3.5}$$

where $\alpha = -E_0\{D_2 | D_1 = 0\} = -\gamma_2 + \sigma_{12}\gamma_1/I(b)$, $\tau^2 = V_0\{D_2 | D_1\} = \sigma_{22} - \sigma_{12}^2/I(b)$, and Φ, ϕ are the standardized normal distribution and density function , respectively. Insertion in (3.2) now gives

$$g_0(b) = (2\pi I(b))^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}\gamma_1^2/I(b)\}(\alpha\Phi(\alpha/\tau) + \tau\phi(\alpha/\tau)), \qquad (3.6)$$

which is exact and hence solves our problem completely, since also its computation is feasible.

To illustrate the general approximation, we shall continue to approximate (3.6) by a simpler expression. We shall do this by expanding e_0 as a function of σ^2 as $\sigma^2 \to 0$, which is mathematically the same asymptotics as obtained by replications of the experiment. Using the expansion $\Phi(-x) \sim (\phi(x)/x)(1 - x^{-2})$ as $x \to +\infty$ we obtain $e_0 \sim \alpha(1 + o(\exp\{-c/\sigma^2\}))$ as $\sigma^2 \to 0$ if $\alpha > 0$ for some constant $c > 0$ depending on α. Since $\alpha = I(\beta_0)$ if $b = \beta_0$, α will by continuity be positive in a neighbourhood of β_0, which is independent of σ^2. Defining the approximation

$$\tilde{e}_0 = \begin{cases} \alpha & \alpha > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (3.7)$$

the relative error is $o(\exp\{-c/\sigma^2\})$ for some $c > 0$ uniformly in $b \in B_0$, where B_0 is some compact neighbourhood of β_0. Since the probability of $\hat{\beta}$ being outside B_0 is also decreasing at exponential rate as $\sigma^2 \to 0$, we shall not worry about that part of the approximation. As a final result we have

$$\tilde{g}_0(b) = h_0(0)\tilde{e}_0, \qquad (3.8)$$

which satisfies

Theorem 3.1. There exists a constant $c > 0$ and a compact neighbourhood B_0 of β_0, such that

$$\tilde{g}_0(b) = g_0(b)(1 + o(\exp\{-c/\sigma^2\})) \qquad (3.9)$$

uniformly in $b \in B_0$ as $\sigma^2 \to 0$, where $\tilde{g}_0$ is given by (3.4), (3.7) and (3.8).

Proof. Follows from Theorem 4.8.

Corollary 3.2. There exists a constant $c > 0$ such that

$$0 \leqq \int_A g_0(b)\,db - \int_A \tilde{g}_0(b)\,db = o(\exp\{-c/\sigma^2\}) \tag{3.10}$$

uniformly in all Borel sets $A \subseteqq B$ as $\sigma^2 \to 0$, where $\tilde{g}_0$ is given by (3.4), (3.7) and (3.8).

Proof. The inequality follows from the trivial fact, that $\tilde{e}_0 \leqq e_0$. The second part follows from Theorem 3.1 and the fact, that $P_0\{\hat{\beta} \notin B_0\} = o(\exp\{-c_1/\sigma^2\})$ for some $c_1 > 0$, which is a concequence of the results in Berk (1972). □

4. Multivariate curved exponential families

We shall now generalize the results of the previous section to multi-dimensional subfamilies of arbitrary exponential families of continuous type. In Section 3, we essentially used the normality via the normality of (D_1,D_2) only. The idea in the general case is instead to use a saddlepoint approximation to the distribution of (D_1,D_2), or rather a simple version of the mixed Edgeworth - saddlepoint approximation, see Barndorff - Nielsen & Cox (1979). The results will, of course, be somewhat more complicated than in the normal case.

Let us first shortly review the saddlepoint approximations; for more thorough accounts on this type of approximations, see Barndorff - Nielsen & Cox (1979), Daniels (1954) and Feller (1971), XVI.7. Let $S = X_1 + \ldots + X_n$ be a sum of n i.i.d. random vectors with density f on $\mathbb{R}^p$, and let f^{*n} denote its n'th convolution with itself, i.e. the density of S. Define

$$g_t(x) = f(x)e^{t'x}/\phi(t), \quad t \in \mathbb{R}^p \tag{4.1}$$

where $\phi(t) = \int e^{t'x} f(x)dx$ is the Laplace-transform of X_1; we shall not at the moment worry about its domain. Now,

$$g_t^{*n}(s) = f^{*n}(s)e^{t's}/\phi(t)^n \tag{4.2}$$

and for a particular s, we may choose $\tilde{t}$, such that $g_{\tilde{t}}$ is 'centered' at s/n, i.e.

$$D \log \phi(\tilde{t}) = \int x g_{\tilde{t}}(x)dx = s/n \; . \tag{4.3}$$

Applying the central limit theorem to $g_{\tilde{t}}^{*n}$, the saddlepoint approximation to $f^{*n}(s)$ follows from (4.2). Letting h_n denote the density of $U = S/n$ this yields

$$h_n(u) = (n/2\pi)^{p/2}|\beta(\tilde{t})|^{-1/2}\exp\{-\tilde{t}'un\}\phi(\tilde{t})^n(1+O(n^{-1})) \qquad (4.4)$$

as $n\to\infty$, where $|\beta(t)|$ is the determinant of

$$\beta(t) = D^2\log\phi(t) = \left(\frac{d^2}{dt_i dt_j}\log\phi(t)\right)_{p\times p}$$

The remaining part of this section deals with the following setup. Let $X_1,\dots,X_n$ be i.i.d. random vectors in $\mathbb{R}^k$, the density of X_1 with respect to some measure μ on $\mathbb{R}^k$ being

$$f(x;\beta) = \exp\{x'\theta(\beta) - \psi(\theta(\beta))\} \qquad (4.5)$$

where $\beta\in B\subseteq\mathbb{R}^p$, B open; $\theta : B\to\mathbb{R}^k$ has a range satisfying

$$\theta(\beta)\in\text{int}\,\Theta = \text{int}\{\theta\in\mathbb{R}^k \mid \psi(\theta) = \int\exp\{x'\theta\}\mu(dx) < \infty\}$$

for all $\beta\in B$. Further, let $X = (X_1,\dots,X_n)$, f_n the density of X and $\bar{X} = \Sigma X_i/n$. As in the previous section we define

$$D_1(\beta) = n^{-1}(D\log f_n(X;\beta))_p$$

$$= (D\theta(\beta))'(\bar{X} - E_\beta\{X_1\})\in\mathbb{R}^p$$

$$D_2(\beta) = n^{-1}(D^2\log f_n(X;\beta))_{p\times p}$$

$$= -I(\beta)/n + (\bar{X} - E_\beta\{X_1\})'_k(D^2\theta(\beta))_{k\times p\times p}$$

where $I(\beta) = n(D\theta(\beta))'_{p\times k}(D^2\psi(\theta(\beta)))_{k\times k}(D\theta(\beta))_{k\times p}$ is the Fisher information matrix. For later reference we shall need the following assumptions:

<u>Assumptions 4.1</u>.

(i) $\bar{X}$ <u>has a continuous density with respect to the Lebesgue</u>

measure on the closed convex support of X_1.

(ii) $\theta : B \to \mathbb{R}^k$ is one-to-one, bicontinouus and three times differentiable on B.

(iii) The Fisher information matrix $I(\beta)$ is regular for all $\beta \in B$.

Let $\beta_0 \in B$ and $b \in B$ be fixed points, and let us again use the convention, that if an argument is omitted, b is understood, whereas a subscript 0 means the value at β_0.

Due to the assumption in Section 3, that only one local maximum of the likelihood function could occur, the limit in (3.2) was equal to the density of $\hat{\beta}$ at b. In general this is not so, but we shall still consider the same quantity

$$\lambda_0(b) = \lim_{\varepsilon \to 0} (\varepsilon^p A_p)^{-1} P_0\{f_n(X;\beta) \text{ has a local maximum within } \|\beta - b\| < \varepsilon\} \tag{4.6}$$

where $A_p = \text{vol}\{\beta \in \mathbb{R}^p \mid \|\beta\| < 1\}$. Thus $\lambda_0(b)$ is, when it exists, the intensity of the point process of local maxima of the likelihood function. Obviously, we have

$$g_0(b) \leqq \lambda_0(b), \quad b \in B \tag{4.7}$$

when the MLE is formally defined as some external point, ∞ say, if the likelihood function has no maximum. The following lemma now generalizes the equality (3.2).

Lemma 4.2. If Assumptions 4.1 are fulfilled, then

$$\lambda_0(b) = h_0(0)e_0, \tag{4.8}$$

where h_0 is the β_0 - density of D_1 and

$$e_0 = E_0\{|-D_2| \cdot I_{\{D_2 \text{ is neg. definite}\}} | D_1 = 0\} \tag{4.9}$$

Proof. See the Appendix.

Remark 4.3. It is clear from the proof, that Lemma 4.2 is not restricted to exponential families, but we have chosen this framework, since in these cases the results are fairly simple and almost no extra conditions are needed to prove their validity.

Our next step is to approximate $h_0(0)$ using the saddlepoint approximation. Since the expectation of D_1 is usually different from zero, the outcome $D_1 = 0$ is a 'large deviation', and the usual normal approximation or the Edgeworth expansions could not be expected to give useful results.

Lemma 4.4. *Let Asumptions 4.1 be fulfilled, and let* $B_0 \subseteq B$ *be any compact neighbourhood of* β_0, *then*

$$h_0(0) = (n/2\pi)^{p/2} |\beta(\tilde{t})|^{-1/2} \exp\{-n[\psi(\theta_0) - \psi(\tilde{\theta}) + \tau(\theta)'(D\theta)\tilde{t}]\}(1 + O(n^{-1})) \quad \text{as } n \to \infty \tag{4.10}$$

uniformly in $b \in B_0$, *where* $\tau(\theta) = D\psi(\theta)$, $\tilde{t}$ *is the unique solution to*

$$(\tau(\tilde{\theta}) - \tau(\theta))'(D\theta) = 0, \quad \tilde{\theta} = \theta_0 + (D\theta)\tilde{t} \in \Theta \tag{4.11}$$

and $\beta(t)$ *is given in (4.14) below.*

Proof. If $n = 1$, the Laplace transform of D_1 and its first two logarithmic derivatives are

$$\phi(t) = E_0\{\exp(t'D_1)\} = \exp\{\psi(\theta_0 + (D\theta)t) - \psi(\theta_0) - \tau(\theta)'(D\theta)t\}, t \in \mathbb{R}^p \tag{4.12}$$

$$D \log \phi(t) = (\tau(\theta_0 + (D\theta)t) - \tau(\theta))'D\theta \tag{4.13}$$

$$D^2 \log \phi(t) = (\beta(t))_{p \times p} = (D\theta)'(D^2\psi(\theta_0 + (D\theta)t))D\theta \tag{4.14}$$

Hence the equation defining the saddlepoint, cf. (4.3), becomes

(4.11). $\tilde{t}$ may be recognized as the MLE in the model $\theta \in \{\theta_0 + (D\theta)t\}$, which is an affine hypothesis in the canonical parameter θ. This fact ensures the existence and uniqueness of a solution to (4.11) satisfying $\tilde{\theta} \in \Theta$, since $\tau(\theta(b))$ belongs to the relative interior of the closed convex support of X_1. The result (4.10) now follows directly from (4.4). □

The approximation to e_0, stated below, is derived from a simple kind of the mixed Edgeworth-saddlepoint approximation, see Barndorff-Nielsen & Cox (1979), expanding the joint distribution of D_1 and D_2 around the same point $\tilde{\theta}$ as used to approximate $h_0(0)$.

Lemma 4.5. Let Assumptions 4.1 be fulfilled and define

$$\gamma_2 = (\tau(\tilde{\theta}) - \tau(\theta))'(D^2\theta) - I(b)/n \tag{4.15}$$

$$\tilde{e}_0 = \begin{cases} |-\gamma_2|, & \text{if } \gamma_2 \text{ is neg. definite} \\ 0 & \text{otherwise} \end{cases} \tag{4.16}$$

then

$$e_0 = \tilde{e}_0(1 + O(n^{-1})) \quad \text{as } n \to \infty \tag{4.17}$$

uniformly on any compact set $B_0 \subseteq B$ on which γ_2 is negatively definite, where e_0 is defined in (4.9).

Let f_θ, $\theta \in \Theta$ denote the θ-density of $\bar{X}$ induced by (4.5) with $\theta(\beta)$ replaced by θ. For fixed d_1 let t_1 be the solution to

$$(\tau(\theta_1) - \tau(\theta(b)))'D\theta(b) = d_1, \quad \theta_1 = (D\theta(b))t_1 + \theta_0 \in \Theta, \tag{4.18}$$

i.e. t_1 is the saddlepoint corresponding to $D_1 = d_1$, when approximating the distribution of D_1. Then, by (4.2)

$$f_{\theta_0}(x) = f_{\theta_1}(x) \exp\{n[\psi(\theta_1) - \psi(\theta_0) - \tau(\theta_1)'(D\theta(b))t_1]\}. \tag{4.19}$$

Since t_1 only depends on x through d_1, so does the entire exponential factor, and since d_1 is an affine function of x, it follows, that the conditional θ_0 - density of X given $D_1 = d_1$ is proportional to $f_{\theta_1}(x)$ on the affine support of X given $D_1 = d_1$. In particular, we may approximate conditional moments, such as e_0, using a normal approximation to f_{θ_1}. Since this approach takes the 'large deviation' event $D_1 = 0$ into account, it is preferable to a direct normal approximation. The result becomes

$$E_0\{D_2 | D_1 = 0\} = E_{\tilde{\theta}}\{D_2 | D_1 = 0\} = \gamma_2 + O(n^{-1}) \qquad (4.20)$$

uniformly in b in any compact set. Since the variance and higher cumulants of D_2 given $D_1 = 0$ are $O(n^{-1})$, while γ_2 is independent of n, (4.17) follows easily. □

<u>Remark 4.6</u>. In some cases it may be possible to evaluate $\int(-d_2)q_{\tilde{\theta}}(0,d_2)d(d_2)$, which would provide a better approximation to e_0. Only in the one-dimensional case, however, would it be feasible to restrict the integration to the set, where $(-d_2)$ is possitively definite. Other improvements are possible, e.g. by including variance-terms in the evaluation of the determinant rather than just computing the determinant of $(-\gamma_2)$.

On combining Lemma 4.2 with Lemma 4.4 and Lemma 4.5, we now have

<u>Corollary 4.7</u>. <u>Let Assumptions 4.1 be fulfilled, and let</u> $B_0 \subseteq B$ <u>be any compact set on which</u> γ_2 <u>is neg. definite, then</u>

$$\lambda_0(b) = \tilde{\lambda}_0(b)(1 + O(n^{-1})) \quad \underline{\text{uniformly in}}\ b \in B_0 \qquad (4.21)$$

<u>where</u>

$$\tilde{\lambda}_0(b) = \begin{cases} (n/2\pi)^{p/2} |\beta(\tilde{t})|^{-1/2} \exp\{-n[\psi(\theta_0) - \psi(\tilde{\theta}) + \tau(\theta)'(D\theta)\tilde{t}]\} |-\gamma_2|, \\ \qquad \underline{\text{if } \gamma_2 \text{ is neg. definite}} & (4.22) \\ 0 \qquad \underline{\text{otherwise}} \end{cases}$$

Proof. Trivial.

It remains now only to be shown, how the approximation to $\lambda_0(b)$ provides an approximation to $g_0(b)$. The answer is simple. Since the probability of a local maximum in a neighbourhood of β_0 not being global tends exponentially to zero, the difference between $g_0(b)$ and $\lambda_0(b)$ tends rapidly to zero, such that $\tilde{\lambda}_0(b)$ also approximates $g_0(b)$.

Theorem 4.8. Let Assumptions 4.1 be fulfilled, then for some neighbourhood B_0 of β_0 and some constant $c > 0$,

$$g_0(b) = \lambda_0(b)(1 + o(\exp\{-cn\}))$$

$$= \tilde{\lambda}_0(b)(1 + O(n^{-1})) \text{ as } n \to \infty \qquad (4.23)$$

uniformly in $b \in B_0$.

Proof. See the Appendix.

Notice, that since B_0 is independent of n, (4.23) is valid for large deviations of the type $\|\sqrt{n}(b - \beta_0)\| = o(\sqrt{n})$ in the normalized variable $\sqrt{n}(\hat{\beta} - \beta_0)$.

Corollary 4.9. Let Assumptions 4.1 be fulfilled, then there exists a constant $c > 0$, such that

$$\int_A g_0(b)\,db - \int_A \tilde{\lambda}_0(b)\,db$$

$$= O(n^{-1}) \int_A g_0(b)\,db + o(\exp\{-cn\}) \text{ as } n \to \infty \qquad (4.24)$$

uniformly in the class of Borel sets A.

Proof. Follows easily from Theorem 4.8 and the fact, that for any neighbourhood B_0 of β_0, a constant $c > 0$ exists, such that $P_0\{\hat{\beta} \notin B_0\} = o(\exp\{-cn\})$ as $n \to \infty$, which follows from the results in Berk (1972). □

Remark 4.10. The advantage of this approximation compared to the normal or Edgeworth approximations is, that apart from the exponentially decreasing term the relative error is $O(n^{-1})$ uniformly in all sets A. This makes the approximation particular useful for calculating tail probabilities.

Remark 4.11. The relative error may be improved from $O(n^{-1})$ to $O(n^{-3/2})$ by a renormalization. There are several ways of doing this; the simplest is to divide $\tilde{\lambda}$ by its integral over B, but this may be infinite. Another method is to adjust $\tilde{\lambda}$ such that at $b = \beta_0$, it equals the value of the third order Edgeworth expansion, i.e. including the $O(n^{-1})$ terms; but the computational work is rather large. A simpler method, which is always valid, is to divide $\tilde{\lambda}_0$ by its integral as approximated by a Gauss-Hermite sum. In the one-dimensional case $(p = 1)$ only 4 terms are required, and the approximation becomes

$$g_0(b) \approx \tilde{\lambda}_0(b)/I,$$

$$I = \sum_{i=1}^{4} w_i \tilde{\lambda}_0(b_i) \exp\{x_i^2\} \sqrt{2/I(\beta_0)} \qquad (4.25)$$

where $b_i = \beta_0 + X_i \sqrt{2/I(\beta_0)}$ and w_i, x_i; $i = 1, \ldots, 4$ may be found in Abramowitz & Stegun (1964), Table 25.10. We shall not prove, that this formula yields a valid asymptotic normalization improving the

$O(n^{-1})$ error to $O(n^{-3/2})$. The proof relies on the fact, that formula (4.25) yields the exact integral, if $\tilde{\lambda}_0$ is a third order Edgeworth expansion.

Remark 4.12. In calculating tail probabilities the best use of the approximation is probably to calculate the tail area rather than its complementary part. The reason is, that essentially the relative error is bounded, such that small probabilities give smaller errors.However, one must convince oneself, that the truncation introduced in (4.16) is of no great importance, This may be indicated by a small density at the boundary, where the truncation becomes effective.

5. An example

To illustrate the performance of the approximation for small n, we shall use an example, chosen not because of practical relevance, but rather as a case, where none of the steps in the approximation are 'too accurate' as in the normal regression, where the saddlepoint approximation (4.10) is exact, and at the same time the computation of the exact density is feasible. The example is one-dimensional, since this makes pictures easier to look at.

Let (Y,Z) be normally distributed with expectations zero and covariance matrix

$$\Sigma = \sigma^2\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad \sigma^2 > 0, \; -1 < \rho < 1$$

and consider the subfamily given by $\sigma^2 = 1$. Defining

$$X = (\tfrac{1}{2}(Y^2 + Z^2), \; YZ)$$

$$\theta(\rho) = (\theta_1, \theta_2) = (-1/(1-\rho^2), \rho/(1-\rho^2))$$

$$\psi(\theta) = -\tfrac{1}{2}\log(\theta_1^2 - \theta_2^2)$$

the model is of the form of Section 4 with ρ being the unknown parameter corresponding to β. Consider n independent replications, then

$$\bar{X} = (\sum_{i=1}^{n}(Y_i^2 + Z_i^2)/2n, \quad \sum_{i=1}^{n} Y_i Z_i/n)$$

and

$$I(\rho) = n(1+\rho^2)/(1-\rho^2)^2.$$

Let us first briefly sketch the derivation of the approximation (4.22), (4.23) to the density $g_\rho(r)$ of $\hat{\rho}$ at $r \in]-1,1[$, when ρ is the true parameter.

The saddlepoint equation (4.11) becomes

$$\tilde{t} - b = a^2 t - 2abt + c \qquad (5.1)$$

where

$$a = r/(1-r^2), \quad b = (2r - \rho(1+r^2))/(1-\rho^2),$$

$$c = -r(1-r^2)/(1-\rho^2). \qquad (5.2)$$

If $r = 0$ then $\tilde{t} = b$, otherwise the saddlepoint is

$$\tilde{t} = (1 + 2ab - (4a^2b^2 + 1 - 4ac)^{1/2})/2a, \quad r \neq 0, \qquad (5.3)$$

for the other solution to (5.1), $\tilde{\theta}$ is outside the range of θ.

A straightforward computation now gives <u>the approximation</u> $\tilde{\lambda}_\rho(r)$ defined in (4.22),

$$\tilde{\lambda}_\rho(0) = \begin{cases} (n/2\pi)^{1/2}(1-\rho^2)^{(n-2)/2}(1-2\rho^2), & |\rho| \leqq 1/\sqrt{2} \\ 0, & |\rho| > 1/\sqrt{2} \end{cases} \qquad (5.4)$$

$$\tilde{\lambda}_\rho(r) = (n/2\pi)^{1/2}\, g_1(\rho,r)/\sqrt{g_2(\rho,r)}\, \exp\{-ng_3(\rho,r)\}, \quad r \neq 0, \qquad (5.5)$$

where

$$g_1(\rho,r) = \begin{cases} (4r^3 + \rho(1-4r^2-r^4))/(r(1-r^2)(1-\rho^2)) + \tilde{t}/(r(1-r^2)), & \text{if positive} \\ 0 & \text{otherwise} \end{cases}$$

$$g_2(\rho,r) = (\tilde{t} - b)(1-r^2)/r + (2(1+r^2)^2 + 8\rho r(\rho r - 1 - r^2))/(1-\rho^2)^2$$

$$g_3(\rho,r) = \tfrac{1}{2}\log[(b-\tilde{t})(1-\rho^2)/(r(1-r^2))] - \tilde{t}r/(1-r^2),$$

and b is given in (5.2), $\tilde{t}$ in (5.3). Although the expression seems complicated, it is quite explicit and quickly calculated on a computer.

The exact density, $g_\rho(r)$, is derived from the Wishart-distribution of $\overline{X}$ on integration along the estimation lines, see e.g. Barndorff-Nielsen (1980), Example 1. The result is

$$f_\rho(r) = \int_0^\infty (r^2 + (1-r^2)s)(s^2 + qs + 1/4)^{(n-2)/2} \exp\{-n(\alpha + \beta s)\}ds$$
$$\cdot\, c_n (1-r^2)^{n-2} / (1-\rho^2)^{n/2}, \qquad (5.6)$$

where

$$q = (1+r^2)/(1-r^2)$$
$$c_n = n^n/(2^{n-2}\Gamma(n/2)^2)$$
$$\alpha = \tfrac{1}{2}(1-r^2)/(1-\rho^2), \qquad \beta = (1+r^2-2r\rho)/(1-\rho^2).$$

If n is even, the integral may be calculated explicitely. In particular, if $n = 2$ we obtain

$$g_\rho(r) = (4/(1-\rho^2))[r^2/(2\beta) + (1-r^2)/(2\beta)^2]\exp\{-2\alpha\}.$$

Note, that here, the computational work increases rapidly with n.

We shall compute the exact and approximate density in three cases, and for comparison also give the usual normal approximation given by

$$\hat{g}_\rho(r) = (n/2\pi)^{1/2}(1+\rho^2)^{1/2}/(1-\rho^2)\exp\{-\tfrac{n}{2}(r-\rho)^2(1+\rho^2)/(1-\rho^2)^2\} \qquad (5.7)$$

We have also calculated the renormalized approximation

$$\tilde{\lambda}_\rho(r) / \int_{-1}^{1} \tilde{\lambda}_\rho(r)dr. \qquad (5.8)$$

The three cases are

I. $\rho = 0$, $n = 10$. Small n, symmetric distribution.

II. $\rho = 0.9$, $n = 10$. Small n, skew distribution.

III. $\rho = 0$, $n = 2$. Extremely small n, symmetric distribution.

The results are given in Fig. 1 - 3 and Table 1 - 3 below. Rather than stating the densities of $\hat{\rho}$ themselves, we have stated the densities of the normalized variable $\sqrt{n}(\hat{\rho} - \rho)$ as functions of $\hat{\rho}$. These are obtained from (5.4), (5.6) - (5.8) on division by $\sqrt{n}$.

Fig. 1. Approximations to the density of $\sqrt{n}(\hat{\rho} - \rho)$ with $\rho = 0$, $n = 10$. Exact density (solid), approximation (5.4) (dashes) and normal approximation (dot-dash).

Fig. 2. Approximations to the density of $\sqrt{n}(\hat{\rho} - \rho)$ with $\rho = 0.9$, n=10. Exact density (solid), approximation (5.4) (dashes) and normal approximation (dot-dash).

Fig. 3. Approximations to the density of $\sqrt{n}(\hat{\rho} - \rho)$ with $\rho = 0$, n=2. Exact density (solid), approximaiton (5.4)(dashes), renormalized approximation (5.8)(dots) and normal approximation (dot-dash).

Table 1. Comparison of approximations to the density of $\sqrt{n}(\hat{\rho} - \rho)$ at $\hat{\rho} = r$, when $\rho = 0$, $n = 10$.

r	exact	normal	approx	renorm
0.00	0.3505	0.3989	0.3989	0.3453
0.10	0.3366	0.3795	0.3853	0.3357
0.20	0.2996	0.3266	0.3477	0.3010
0.35	0.2244	0.2162	0.2645	0.2290
0.50	0.1476	0.1143	0.1720	0.1489
0.70	$5.355 \cdot 10^{-2}$	$3.443 \cdot 10^{-2}$	$5.917 \cdot 10^{-2}$	$5.122 \cdot 10^{-2}$
0.90	$2.176 \cdot 10^{-3}$	$6.951 \cdot 10^{-3}$	$2.270 \cdot 10^{-3}$	$1.965 \cdot 10^{-3}$
0.99	$3.859 \cdot 10^{-7}$	$2.969 \cdot 10^{-3}$	$3.933 \cdot 10^{-7}$	$3.405 \cdot 10^{-7}$

Table 2. Comparison of approximations to the density of $\sqrt{n}(\hat{\rho} - \rho)$ at $\hat{\rho} = r$, when $\rho = 0.9$, $n = 10$.

r	exact	normal	approx	renorm
0.50	$1.284 \cdot 10^{-5}$	$1.074 \cdot 10^{-17}$	$6.950 \cdot 10^{-6}$	$6.893 \cdot 10^{-6}$
0.70	$1.088 \cdot 10^{-2}$	$1.247 \cdot 10^{-4}$	$8.547 \cdot 10^{-3}$	$8.477 \cdot 10^{-3}$
0.80	0.2983	0.2303	0.2761	0.2738
0.85	1.152	1.509	1.130	1.121
0.88	2.118	2.555	2.130	2.113
0.90	2.777	2.825	2.825	2.802
0.92	3.089	2.555	3.167	3.141
0.95	2.109	1.509	2.172	2.154
0.99	$2.494 \cdot 10^{-2}$	$3.708 \cdot 10^{-1}$	$2.547 \cdot 10^{-2}$	$2.526 \cdot 10^{-2}$

Table 3. Comparison of approximations to the density of $\sqrt{n}(\hat{\rho} - \rho)$ at $\hat{\rho} = r$, when $\rho = 0$, $n = 2$.

r	exact	normal	approx	renorm
0.00	0.2601	0.3989	0.3989	0.2813
0.20	0.2611	0.3833	0.4069	0.2869
0.40	0.2748	0.3400	0.4340	0.3060
0.60	0.3264	0.2783	0.4950	0.3491
0.80	0.4511	0.2104	0.6041	0.4260
0.95	0.6258	0.1618	0.7208	0.5083

In case I ($\rho = 0$, $n = 10$), it is seen (Fig. 1, Table 1), that except for the extreme tail, the normal approximation does quite well. The approximation (5.4) is slightly worse in the main part of the distribution, but keeps the shape better leading to an excellent renormalized approximation, which has not been drawn since it is hardly distinguishable from the exact density. Note, that the approximation (5.4), renormalized or not keeps its degree of approximation throughout the range. No truncation (see (4.22) and (5.4)) is needed, when $\rho = 0$.

In case II ($\rho = 0.9$, $n = 10$) the distribution is skew, and the normal approximation is useless. The approximation (5.4), however, does quite well throughout the range. Here, the effect of renormalization is vanishing, since the integral of the approximation is 1.008. In this case the approximation is truncated at 0.35324 at a density of approx. 10^{-7}.

In case III ($\rho = 0$, $n = 2$), n is so small, that hardly any approximation can be expected to work. Both (5.4) and the normal approximation are numerically useless as direct approximations (Fig. 3, Table 3), but the approximation (5.4) again has the right shape, and its renormalized verion does surprisingly well.

The comparison with the normal approximation has only been included to give an impression of the magnitude of the deviations. In a thorough investigation of the behaviour it would be more relevant to compare with the second-order Edgeworth expansion, which has an error of order $O(n^{-1})$, and which better approximates skew distributions. Its tail behaviour is, however, not in general better than that of the normal distribution.

It seems, that the approximation (5.4) behaves similar to the saddle-point approximation on which it is based, see Daniels (1954); namely keeping its relative error fairly constant throughout the range, such that the renormalized approximation works extremely well. When calculating tail probabilities of magnitude 0.01, say, a relative error of 50% is often of no great importance, and the approximation may safely be used directly, unless the truncation is of importance.

6. The multi-dimensional normal regression model

In this section, we shall specialize the results of Section 4 to obtain a simple explicit approximation to the density of the MLE of the (multivariate) parameter in the important class of nonlinear normal regression models. As in Section 3 we shall consider the asymptotics obtained by letting the variance tend to zero, which is equivalent (mathematically) to simple replications.

Let $X = (X_1, \ldots, X_k)$ be normally distributed with expectation vector $\mu(\beta) = (\mu_1(\beta), \ldots, \mu_k(\beta))$ and covariance matrix $\Sigma = \sigma^2 I_{k \times k}$, where $\beta \in B \subseteq \mathbb{R}^p$ is the unknown parameter, B is open, σ^2 is considered to be known, and $\mu : B \to \mathbb{R}^k$ is a known function.

As previously, $\hat{\beta}$ is the MLE of β; $\beta_0 \in B$ and $b \in B$ are arbitrary fixed points, and

$$I(\beta) = (D_\mu(\beta))' D_\mu(\beta) / \sigma^2 \tag{6.1}$$

is the Fisher information matrix. Since the density of X is of the form

$$f(X;\beta) = c(\beta) \cdot \{\exp \mu(\beta)' X / \sigma^2\}$$

this is a curved exponential family model as discussed in Section 4. The Assumptions 4.1 are equivalent to

Assumptions 6.1.

(i) $\mu : B \to \mathbb{R}^k$ is one-to-one, bicontinuous and three times differentiable.

(ii) The Fisher information matrix $I(\beta)$ is regular for all $\beta \in B$.

A direct computation, either by insertion in the results of Section 4 or using the normality of (D_1, D_2) in combination with Lemma 4.2, now shows, that the approximation (4.22) becomes

$$\tilde{\lambda}_0(b) = (2\pi)^{-p/2}|I(b)|^{-1/2}\tilde{e}_0 \cdot$$

$$\exp\{-\tfrac{1}{2}(\mu(\beta_0) - \mu(b))'D\mu(b)I(b)^{-1}D\mu(b)'(\mu(\beta_0) - \mu(b))/\sigma^4\}$$

$$= (2\pi)^{-p/2}|I(b)|^{-1/2}\,\tilde{e}_0 \cdot \exp\{-\frac{1}{2\sigma^2}\,||P_b(\mu(\beta_0) - \mu(b))||^2\}, \qquad (6.2)$$

where P_b is the projection matrix onto the subspace spanned by the columns of $D\mu(b)$, and

$$\tilde{e}_0 = \begin{cases} |-\gamma_2| & \text{if } \gamma_2 \text{ is negatively definite} \\ 0 & \text{otherwise} \end{cases} \qquad (6.3)$$

$$\gamma_2 = \frac{1}{\sigma^2}(\mu(\beta_0) - \mu(b))'(I_{k\times k} - D\mu(b)(\sigma^2 I(b))^{-1}D\mu(b)')D^2\mu(b) - I(b)$$

$$= \frac{1}{\sigma^2}(\mu(\beta_0) - \mu(b))'(I_{k\times k} - P_b)D^2\mu(b) - I(b) \qquad (6.4)$$

<u>Remark 6.2</u>.

The asymptotic behaviour of this approximation is stated in Theorem 4.8 and Corollary 4.9. There are, however, less approximations involved in this case, since the approximation (4.10) to $h_0(0)$ here is exact, and also γ_2 is exactly equal to $E_0\{D_2|D_1 = 0\}$. Both of these approximations contributed with a relative error of order $O(n^{-1})$ in the general case. Hence, it might be worthwhile also to remove the last $O(n^{-1})$ - error by including the variance terms in the approximation to e_0. Even the higher-order terms (in powers of n^{-1}) may be removed in this way, leaving only exponentially decreasing errors, but if the dimension p is large, this requires quite a lot of computation. We shall not state these refined expansions, which are easily written down for a specific p.

7. Appendix

Proof of Lemma 4.2. The event $M_1(\varepsilon)$, say, that $f(X;\beta)$ has a local maximum at some β in $\{\|\beta - b\| < \varepsilon\}$ is the same as the event, that $D_1(\beta) = 0$ and $D_2(\beta)$ is negatively definite for some β in the same set, except if $D_1(\beta) = 0$ and $D_2(\beta) = 0$, which may be disregarded because of Assumption 4.1(i). We shall show, that furthermore $M_1(\varepsilon)$ may be replaced by the event

$$M_2(\varepsilon) = \{S_1(\beta) = 0 \text{ for some } \beta \text{ in } \{\|\beta - b\| < \varepsilon\}, \text{ and } D_2(b) \text{ is neg. definite}\}$$

where $S_1(\beta) = D_1(b) + D_2(b)(\beta - b) = D_1 + D_2(\beta - b)$ is the linear approximation to $D_1(\beta)$ around $\beta = b$.

Observe, that since the Laplace transform $E_0\{e^{s'X}\}$ exists in a neighbourhood of zero, there exists a $K_1 > 0$, such that

$$P_0\{\|\bar{X} - \tau(\theta)\| \geqq K_1 \log \varepsilon^{-1}\} = o(\varepsilon^p) \quad \text{as } \varepsilon \to 0 \tag{7.1}$$

and for some positive constants K_2, K_3 we have

$$\|D_1(\beta) - S_1(\beta)\| \leqq K_2\ \varepsilon^2 \log \varepsilon^{-1} \tag{7.2}$$

$$\|D_2(b)\| \leqq K_3 \log \varepsilon^{-1} \tag{7.3}$$

$$\text{if } \|\bar{X} - \tau(\theta)\| < K_1 \log \varepsilon^{-1}.$$

Now, let $0 < \delta < 1$ be fixed, then

$$(1+\delta)^p \lim_{\varepsilon \to 0} (\varepsilon^p A_p)^{-1} P_0(M_1(\varepsilon))$$

$$= (1+\delta)^p \lim_{\varepsilon \to 0} (\varepsilon^p (1+\delta)^p A_p)^{-1} P_0(M_1(\varepsilon(1+\delta)))$$

$$\geqq \lim_{\varepsilon \to 0} (\varepsilon^p A_p)^{-1} [P_0(M_2(\varepsilon)) - P_0(M_2(\varepsilon) \setminus M_1(\varepsilon(1+\delta)))]. \tag{7.4}$$

But, if $S_1(\beta_1) = 0$, $||\beta_1 - b|| < \varepsilon$, and $||\overline{X} - \tau(\theta)|| < K_1 \log \varepsilon^{-1}$, then

$$\{S_1(\beta) \mid ||\beta - \beta_1|| < \delta\varepsilon\} \supseteqq \{y \in \mathbb{R}^p \mid ||y|| < \lambda\delta\varepsilon\}$$

where λ is the smallest eigenvalue of D_2, such that by (7.2),

$$\{D_1(\beta) \mid ||\beta - \beta_1|| < \delta\varepsilon\} \supseteqq \{y \in \mathbb{R}^p \mid ||y|| < \lambda\delta\varepsilon - K_2\varepsilon^2 \log \varepsilon^{-1}\},$$

which contains zero if $\lambda > K_2\varepsilon \log \varepsilon^{-1}/\delta$. Since also $||D_1|| \leq ||D_2|| \; ||\beta_1 - b|| \leqq K_3\varepsilon \log \varepsilon^{-1}$, we obtain

$$P_0(M_2(\varepsilon) \setminus M_1(\varepsilon(1+\delta))) \leqq P_0\{||\overline{X} - \tau(\theta)|| \geqq K_1 \log \varepsilon^{-1}\}$$

$$+ P_0\{||D_1|| \leqq K_3\varepsilon \log \varepsilon^{-1}\} \cap \{\lambda \leqq K_2\varepsilon \log \varepsilon^{-1}/\delta\}$$

$$= o(\varepsilon^p) + O((K_3\varepsilon \log \varepsilon^{-1})^p \cdot (K_2\varepsilon \log \varepsilon^{-1}/\delta)) = o(\varepsilon^p)$$

as $\varepsilon \to 0$, proving that

$$\lim_{\varepsilon\to 0}(\varepsilon^p A_p)^{-1} P_0(M_1(\varepsilon)) \geqq \lim_{\varepsilon\to 0}(\varepsilon^p A_p)^{-1} P_0(M_2(\varepsilon)), \tag{7.5}$$

since δ was arbitrary. The other inequality follows similarly.

By Assumptions 4.1, we may write $D_2 = A(D_1) + Y - I(b)$, such that A is a linear function and (D_1, Y) has a continuous density $\zeta_0(d_1, y)$, say, on its closed convex support, which contains $(0, E_0\{Y \mid D_1 = 0\})$ as an interior point. Thus by continuity, we have

$$\lambda_0(b) = \lim_{\varepsilon\to 0} (\varepsilon^p A_p)^{-1} P_0(M_2(\varepsilon))$$

$$= \lim_{\varepsilon\to 0} (\varepsilon^p A_p)^{-1} \int_{d_2 \text{ neg.def.}} \int_{d_1 \in d_2(B_\varepsilon)} \zeta_0(d_1, y)\, d(d_1)\, dy$$

$$= \int_{y - I(b) \text{ neg.def.}} |I(b) - y| \zeta_0(0, y)\, dy = h_0(0) e_0$$

where $B_\varepsilon = \{x \in \mathbb{R}^p \mid ||x|| < \varepsilon\}$, and h_0 and e_0 are defined in the Lemma. □

Proof of Theorem 4.8. If $\overline{X} = \tau(\theta_0)$, then the likelihood function has a global maximum at β_0. Hence, by continuity, if $D_1(\beta_0) = 0$ and $\overline{X}$ lies within a certain neighbourhood of $\tau(\theta_0)$, the local maximum at β_0 will also be global. By another continuity argument, using Assumption 4.1 (ii), there is a neighbourhood B_0' of β_0, and for each $b \in B_0'$ a neighbourhood $T(b)$, such that if $D_1(b) = 0$ and $\overline{X} \in T(b)$, then the likelihood function has a global maximum at b. Now, for some constant c(b),

$$P_0\{\overline{X} \notin T(b) \mid D_1(b)=0\} = o(\exp\{-c(b)n\}) \quad \text{as } n \to \infty.$$

This proves the first equality of (4.23), since also the uniformity follows by continuity. The second equality follows by Corollary 4.7. □

References

Abramowitz, M. & Stegun, I.A. (ed.) (1964). Handbook of Mathematical Functions. National Bureau of Standards, Washington, D.C.

Barndorff-Nielsen, O. (1980) Conditionality resolutions. Biometrika 67, 293-310.

Barndorff-Nielsen O. & Cox, D.R. (1979). Edgeworth and saddlepoint approximations with statistical applications J.R. Statist. Soc. B 41, 279-312.

Berk, R.H. (1972). Constistency and asymptotic normality of mle's for exponential models. Ann. Math. Statist. 43, 193 - 204.

Daniels, H.E. (1954). Saddlepoint approximation in statistics. Ann. Math. Statist. 25, 631-50.

Feller, W. (1971). An Introduction to Probability Theory and its Applications. Vol. II. Wiley, New York.

PREPRINTS 1980

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Olkin, Ingram and Væth, Michael: Maximum Likelihood Estimation in a Two-way Analysis of Variance with Correlated Errors in One Classification.

No. 2 Skovgaard, Ib M.: Transformation of an Edgeworth Expansion by a Sequence of Smooth Functions.

No. 3 Asmussen, Søren: Equilibrium Properties of the M/G/1 Queue.

No. 4 Johansen, Søren and Keiding, Susanne: A Family of Models for the Elimination of Substrate in the Liver.

No. 5 Skovgaard, Ib M.: Edgeworth Expansions of the Distributions of Maximum Likelihood Estimators.

No. 6 Tjur, Tue: A Connection Between Rasch's Item Analysis Model and a Multiplicative Poisson Model.

No. 7 Lauritzen, Steffen L. : Examples of Extreme Point Models in Statistics.

PREPRINTS 1981

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Johansen, Søren: Asymptotic Inference in Random Coefficient Regression Models.

No. 2 Asmussen, Søren: On the Role of a Certain Eigenvalue in Estimating the Growth Rate of a Branching Process.

No. 3 Lauritzen, Steffen.: Time Series Analysis in 1880. A Discussion of Contributions made by T.N. Thiele.

No. 4 Asmussen, Søren: Conditioned Limit Theorems Relating a Random Walk to its Associate, with Applications to Risk Reserve Processes and the GI/G/1 Queue.

No. 5 Johansen, Søren: The Statistical Analysis of a Markov Branching Process.

No. 6 Asmussen, Søren: Time - Dependent Approximations in some Queueing Systems with Imbedded Markov Chains Related to Random Walks.

No. 7 Skovgaard, Ib M.: A Second - Order Investigation of Asymptotic Ancillarity.

No. 8 Rootzén, Holger: The Rate of Extremes of Stationary Normal Sequences.

No. 9 Skovgaard, Ib M.: Large Deviation Approximations for Maximum Likelihood Estimators.