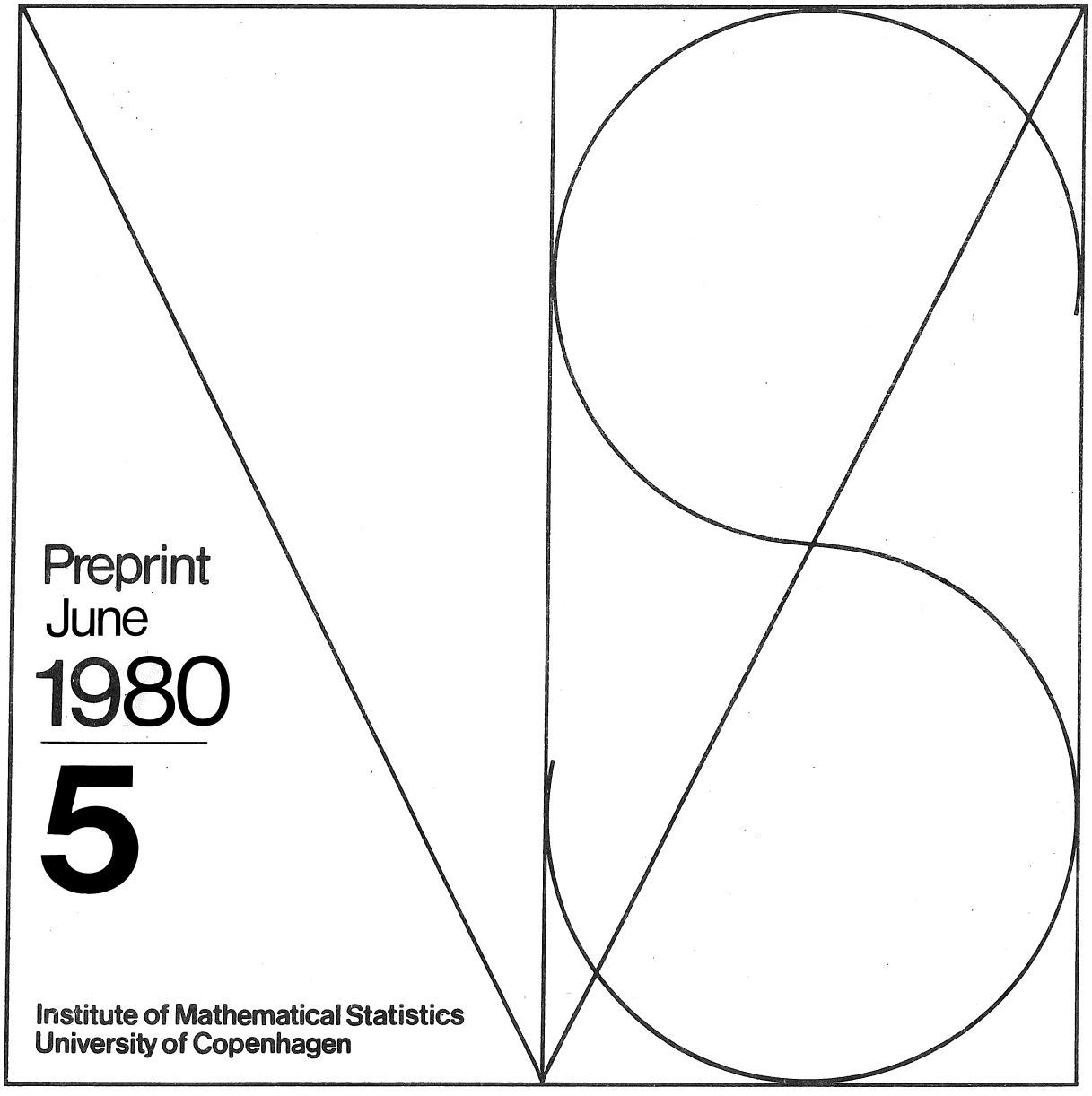
Ib M. Skovgaard

Edgeworth Expansions of the Distributions of Maximum Likelihood Estimators



Preprint
June
1980

5

Institute of Mathematical Statistics
University of Copenhagen

Ib M. Skovgaard

EDGEWORTH EXPANSIONS OF THE DISTRIBUTIONS OF MAXIMUM LIKELIHOOD ESTIMATORS*

Preprint No. 5

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

June 1980

ABSTRACT. In this paper we use the method described in Skovgaard (1980) to derive Edgeworth expansions of the distributions of maximum likelihood estimators in the general (non i.i.d.) case. Comparatively simple sufficient conditions for the validity of the expansion are derived and further simplification obtained in the non-linear normal regression models. A precise description of how to compute the expansion is given, and the first four terms of the corresponding stochastic expansion are given in an explicite form. In case of a smooth hypothesis of an exponential family, also an explicite version of the approximate cumulants, needed to compute the first four terms of the Edgeworth expansion, is given. It is shown that corresponding results for functions of the estimator are easily derived from the original expansion.

Key words: approximate cumulants, Edgeworth expansion, exponential family models, maximum likelihood estimator, non-linear normal regression.

1. Introduction.

The purpose of the present paper is to derive simple sufficient conditions for the validity of an Edgeworth expansion of the maximum likelihood estimator in the non-i.i.d. case, and to compute the quantities needed for this expansion. The expansion is obtained by formally calculating moments of a stochastic expansion which is a Taylor series expansion of the maximum likelihood estimator.

The proof of the main theorem is based on the results of Skovgaard (1980), the method being similar to the one used in Bhattacharya & Ghosh (1978). In the i.i.d. case the moments used in the expansions may be identified with those given in a number of papers, e.g. Shenton & Bowman (1977).

In Section 2 we present the notation and the regularity conditions used in this paper. Section 3 contains the main results as Theorem 3.5 and Corollary 3.10 proving the validity of Edgeworth expansions of the maximum likelihood estimator and functions of this. Also a method of obtaining the stochastic expansion of the MLE is described. In Section 4 we relate these results to smooth hypothesis in exponential families, since in this (very common) case, the cumulants can be given in a more explicit form.

In Section 5 we condider an important class of models, namely the non-linear regression model with normally distributed errors. Here the conditions of Theorem 3.5 may be replaced by one simple condition, and the results also simplifies considerably.

In Section 6 we state in an explicit form the first four terms of the stochastic expansion in the general case, some of the approximate cumulants needed for the Edgeworth expansion in the case of an exponential family model and in the case of the non-linear normal regression model.

2. Notation and basicassumptions.

Let V, W be finite dimensional Euclidean vector spaces. If v_1, v_2 belongs to V, then $< v_1, v_2 >$ denotes their inner product and

$$|| v_1 || = < v_1, v_1 >^{\frac{1}{2}} \tag{2.1}$$

$$(v_1{}'^m) = (v_1,\ldots,v_1) \in V^m, \; m \in \mathbb{N} \tag{2.2}$$

$\mathcal{B}(V)$ is the Borel system on V and $\mathcal{B}_k$ that on $\mathbb{R}^k$. $\mathrm{Hom}(V,W)$, the class of linear mappings of V into W, and $B_p(V,W)$, the class of p-linear, symmetric mappings of V^p into W, are in the natural way given the structure of Euclidean vector spaces, e.g. the norm of $A \in B_p(V,W)$ is given by

$$|| A || = \sup\{ || A(v'^p) || \;\Big|\; || v || \leq 1 \} \tag{2.3}$$

We shall use the usual isomorphisms between vector spaces, e.g. $B_p(V,W) \simeq B_{p-1}(V, \mathrm{Hom}(V,W))$ without distinguishing between these. If $A \in \mathrm{Hom}(V,W)$, then A^* denotes its adjoint, i.e.

$$< A(v), w > = < v, A^*(w) > , \; v \in V, \; w \in W. \tag{2.4}$$

$C^p(V,W)$ denotes the class of p times continuously differentiable functions of V into W. The p'th differential of $f \in C^p(V,W)$ at v_0 is the function in $B_p(V,W)$ given by

$$D^p f(v_0)(v'^p) = \frac{d^p}{dh^p} f(v_0 + hv)\Big|_{h=0} , \; v, v_0 \in V, h \in \mathbb{R} \tag{2.5}$$

Cumulants and moments of a distribution on V will, if they exist, be considered as multilinear, symmetric forms on V, e.g. the p'th moment, μ_p, of a random vector X in V is given by

$$\mu_p(v'^p) = E\{< v, X >^p\} \ , \ v \in V \qquad (2.6)$$

where $E\{...\}$ denotes expectation.

The normal density on V with mean zero and variance equal to the inner product on V will be denoted ϕ (V is understood), and $\phi_{\mu,\Sigma}$ will denote the normal density with mean μ und variance Σ.

The Cramér-Edgeworth polynomials $(\tilde{P}_r)$ are as usual defined by the formal identity

$$\sum_{r=0}^{\infty} u^r \tilde{P}_r(v:\{\chi_j\}) = \exp\{\sum_{r=1}^{\infty} u^r \chi_{r+2}(v'^{r+2})/(r+2)!\} \qquad (2.7)$$

where $\{\chi_j\}$, $j \in \mathbb{N}$ are the cumulants of a distribution. Also, if $\Sigma \in B_2(V,\mathbb{R})$ is regular $P_r(-\phi_{\mu,\Sigma} :\{\chi_j\})$ is the density of the finite signed measure with characteristic function $\tilde{P}_r(iv : \{\chi_j\})\exp\{-\frac{1}{2}\Sigma(v , v) + i\, v\mu\}$ obtained by formally substituting the differential operator for $(-iv)$ in $\tilde{P}_r(iv : \{\chi_j\})$ and using this on $\phi_{\mu,\Sigma}$. In particular $P_r(-\phi_{0,\Sigma}:\{\chi_j\})(v)$ is a polynomial in $v \in V$ multiplied by $\phi_{0,\Sigma}(v)$.

The order symbols o and O are unless otherwise stated used in the sense "as $n \to \infty$".

Let $(E,\mathcal{E})$ be a measurable space, and (P_θ), $\theta \in \Theta \subseteq V$ a family of probability measures on $(E,\mathcal{E})$ <u>dominated</u> by a measure μ. V is a finite dimensional Euclidean vector space and Θ is open in V. Denine

$$f(x;\theta) = (dP_\theta/d_\mu)(x) \ , \ x \in E \ , \ \theta \in \Theta \qquad (2.8)$$

as some version of the Radon-Nikodym derivative of P_θ with

respect to μ. Throughout the paper the following regularity conditions will be assumed to hold.

Assumptions 2.1. A version $f(.;.)$ of the Radon-Nikodym derivatives (2.1) exists, such that for each fixed $\theta_0 \in \Theta$ the following conditions hold. Define

$$E_0 = \{x \in E \mid f(x;\theta_0) > 0\} \tag{2.9}$$

Then for some integer $p \geq 2$,

I. $f(x;\theta)$ is p times continuously differentiable at θ_0 with respect to θ for all $x \in E_0$

II. $$E\{\| D^j \log f(X;\theta_0) \|^2\} < +\infty, \quad 1 \leq j \leq p \tag{2.10}$$

III. $$E\{D f(X;\theta_0)\} = E\{D^2 f(X;\theta_0)\} = 0 \tag{2.11}$$

IV. $V\{D \log f(X;\theta_0)$ is regular (2.12)

Here and in the sequel X is a random variable having distribution P_{θ_0}, $E\{\cdot\}$ and $V\{\cdot\}$ denote expectation and variance in this distribution.

Remark 2.2. Condition (2.11) is the identity obtained by differentiating the integral of the density with respect to θ. The assumption that this may be done inside the integral holds, if $\| D f \|$ and $\| D^2 f \|$ are bounded on $E \times U(\theta_0)$ by functions independent of θ and with finite expectations (w.r.t. P_{θ_0}), where $U(\theta_0)$ is a neighbourhood of θ_0.

Neither (2.11) nor (2.12) are necessary assumptions, but they are

assumed to avoid technical problems. Notice, that (2.10) makes sence because $D^j \log f(X;\theta_0)$ is defined with probability one. Define

$$E_j = E\{D^j \log f(X;\theta_0)\} \in B_j(V,\mathbb{R}) \tag{2.13}$$

$$S_j = D^j \log f(X;\theta_0) - E_j \in B_j(V,\mathbb{R}) \tag{2.14}$$

$$\Sigma_j = V\{S_j\} \in B_2(B_j(V,\mathbb{R}),\mathbb{R}) \tag{2.15}$$

By (2.11) we have

$$E_1 = 0 \ , \ \Sigma_1 = - E_2 \in B_2(V,\mathbb{R}) \simeq \mathrm{Hom}(V,V) \tag{2.16}$$

3. Main results.

In this section we consider a sequence of experiments indexed by $N \in \mathbb{N}$, each setup of the form introduced in Section 2. All the quantities used except the parameter thus depend on n, but for notational simplicity we shall not always wrete the index n. The index $n \in \mathbb{N}$ may be replaced by any $i \in I$, where I is a set directed to the right, indexing a system of statistical fields with the same parameter space Θ. The purpose of this section is to derive an Edgeworth expansion of the distribution of the maximum likelihood estimator (MLE) of $\theta \in \Theta$, based on the assumption, that the first p derivatives of the logarithm of the likelihood function at θ_0 may be approximated in distribution by an Edgeworth series when θ_0 is the true value of the parameter. The notation used is coordinate-free, but the main results of the paper are summarized in terms of coordinates in Section 6.

Assumptions 3.1. Integers $s \geq 2$ and $m \in \mathbb{N}$ exist, such that $(S_1,\ldots,S_p)$ has absolute moment of order s, and a sequence of linear mappings $A_n: B_1(V,\mathbb{R})x\ldots x\, B_p(V,\mathbb{R}) \to \mathbb{R}^m$ exists, satisfying

I. $$V\{A_n(S_1,\ldots,S_p)\} = 1_{\mathbb{R}^m} \tag{3.1}$$

II. $$P\{(S_1,\ldots,S_p) \in B\} = \int_{A_n(B)} \xi_n(t)\,dt + o(\beta_n) \tag{3.2}$$

uniformly in the system of Borel-subsets of the linear space spanned by $(S_1,\ldots,S_p)$, where

$$\xi_n(t) = \left(\sum_{r=0}^{S-2} P_r(-\phi:\{\chi_\nu\})\right)(t) \tag{3.3}$$

χ_ν = ν'th cumulant of $A_n(S_1,\ldots,S_p)$, $2 \leq \nu \leq s$

$$\beta_n = \begin{cases} (\sup\{\|\chi_\nu\|^{1/(\nu-2)} \mid 3 \leq \nu \leq s\})^{s-2} & \text{if } s \geq 3 \\ 1 & \text{if } s = 2 \end{cases} \tag{3.4}$$

Assumptions. 3.2. An $\alpha > 0$ and a sequence (λ_n), $n \in \mathbb{N}$, of positive real numbers exist, such that

I. $$\lambda_n^{p-1} = o(\beta_n) \tag{3.5}$$

II. $$\|E_{j+1}\,o(\Sigma_1^{-\frac{1}{2}})^{,j+1}\|/j\,! = O(\lambda_n^{j-1}) \ , \ 2 \leq j \leq p-1 \tag{3.6}$$

III. $$\|\Sigma_j\,o(\Sigma_1^{-\frac{1}{2}})^{,2j}\|^{\frac{1}{2}}/(j-1)\,! = O(\lambda_n^{j-1}) \ , \ 2 \leq j \leq p \tag{3.7}$$

IV. A sequence (D_n), $n \in \mathbb{N}$ of sets exists, such that $P\{X_n \in D_n\} = 1 - o(\beta_n)$, and for all $x \in D_n$, $\theta \in \theta_0 + \Sigma_1^{-\frac{1}{2}}(H_n(\alpha))$, $D^p \log f(x;\theta)$ exists, and

$$\sup\{\| (D^p \log f(x;\theta) - (S_p+E_p)) \circ (\Sigma_1^{-\frac{1}{2}})', p \| \;\Big|\; \theta \in \theta_0 + \Sigma_1^{-\frac{1}{2}}(H_n(\alpha))\} = o(\beta_n) \tag{3.8}$$

where

$$H_n(\alpha) = \{z \in V \;\Big|\; \| z \| \leq \rho_n(\alpha)\}$$

$$\rho_n(\alpha) = ((2+\alpha)\log \beta_n^{-1})^{\frac{1}{2}} \tag{3.9}$$

Remark 3.3. Assumption 3.1 assure that an Edgeworth expansion of the distribution of $(S_1,\ldots,S_p)$ is available, and Assumptions 3.2 that the derivatives of the MLE of θ w.r.t. $(S_1,\ldots,S_p)$ are sufficiently well behaved.

Remark 3.4. Notice, that since Σ_1 is regular, we have

$$\| E_{j+1} \circ (\Sigma_1^{-\frac{1}{2}})', j+1 \| = \sup\{ | E_{j+1}(\Sigma_1^{-\frac{1}{2}}(u)', j+1) | \;\Big|\; u \in V\}$$

$$= \sup\{ | E_{j+1}(v', j+1) | / \Sigma_1(v,v)^{(j+1)/2} \mid v \in V\}$$

such that (3.6) and (3.7) are not as hard to prove as it may seem; see the exponential regression example in Section 5.

As in Skovgaard (1980) we define the formal cumulants (and formal moments) of polynomials of $(S_1,\ldots,S_p)$ as the cumulants (and moments) computed in the usual way in terms of the cumulants of $(S_1,\ldots,S_p)$, except that the cumulants of $(S_1,\ldots,S_p)$ of order higher than s are defined as zero.

Theorem 3.5. Suppose Assumptions 2.1, 3.1 and 3.2 hold. Then a sequence $\hat{\theta}_n$ of estimators of θ exists, such that with probability $1 - o(\beta_n)$, $\hat{\theta}_n$ is a unique maximum of the likelihood function in

the interior of $\theta_0 + \Sigma_1^{-\frac{1}{2}} H_n(\alpha)$, and the following expansion holds

$$P\{\hat{\theta}_n - \theta_0 \in B\}$$

$$= \int_{B-\kappa_1} \eta_n(t)\,dt + o(\beta_n) \quad \text{uniformly in } B \in \mathcal{B}(V) \tag{3.10}$$

where $B - \kappa_1 = \{t \mid t + \kappa_1 \in B\}$ and

$$\eta_n(t) = \sum_{r=0}^{q-2} P_r(-\phi_{0,\kappa_2} : \{\kappa_\nu\})(t) \tag{3.11}$$

$$q = \max\{p,s\} \tag{3.12}$$

and $\{\kappa_\nu\}$ are approximate cumulants of the polynomial

$$Y_1 + \sum_{j=2}^{p-1} A_j(Y_1,\ldots,Y_j) \ , \ Y_j = \Sigma_1^{-1} S_j \tag{3.13}$$

where A_j is a homogeneous polynomial of degree j in $(Y_1,\ldots,Y_j)$ computed as described in Remark 5.6, and $\{\kappa_\nu\}$, $1 \le \nu \le q$ are computed as described in Remark 5.7. A_1 through A_4 are given in Section 6.

Remark 5.6. Computation of the A's.

Consider the Taylor-series expansion of the likelihood equation

$$\Sigma_1^{-1} S_1 \approx - \sum_{j=2}^{p} \Sigma_1^{-1} (E_j + S_j)(\hat{\theta}_n - \theta_0)^{,j-1} / (j-1)! \tag{3.14}$$

Considering $(S_2,\ldots,S_p)$ as fixed the derivatives of $(\hat{\theta}_n - \theta_0)$ with respect to S_1 at $S_1 = 0$ may be expressed in terms of the derivatives of S_1 with respect to $(\hat{\theta}_n - \theta_0)$ at zero. These (former) derivatives are easily derived recursively and it is seen, that they are polynomials in $((1_V - Y_2)^{-1}, Y_3,\ldots,Y_p)$. Expanding

$$(1_V - Y_2)^{-1} = \sum_{j=0}^{N} (Y_2)^k + o(\| Y_2 \|^N) \text{ as } \| Y_2 \| \to 0 \quad (3.15)$$

we obtain an expansion of $(\hat{\theta}_n - \theta_0)$ as a polynomial in the Y's around $(Y_1, Y_2) = (0,0)$. In particular the Taylor series expansion of $(\hat{\theta}_n - \theta_0)$ with respect to $(Y_1,\ldots,Y_p)$ around $(0,\ldots,0)$ is obtained as (3.13) by equating $A_j(Y_1,\ldots,Y_p)$ to the sum of the terms of power j in $(Y_1,\ldots,Y_p)$. It is easy to see that A_j only depends on $(Y_1,\ldots,Y_j)$.

Remark 3.7. Computation of the κ's.

By the results of Leonov & Shiryaev (1959) and Skovgaard (1980) it follows, that the approximate cumulants (κ_ν), $1 \leq \nu \leq q$ may be calculated as follows. Recall, that $q = \max\{p,s\}$.

To calculate κ_ν, $1 \leq \nu \leq q$, raise (3.13) to the power ν and consider each term, omitting terms of power greater than $\nu + p - 2$ in the Y's, and also of power $\nu + p - 2$ if this is odd. For each of the remaining terms compute its mean in terms of the cumulants of the Y's, and omit terms for which

I. The "partition" corresponding to the cumulants is decomposable; see Leonov & Shiryaev (1959) or for a short description Brillinger (1975).

II. The number of cumulants entering the term is strictly less than $x - (\nu+q-2)/2$, where x is the degree (in Y) of the term.

κ_ν is then obtained as the sum of the remaining terms.

Using this method, κ_ν may be written down almost immediately from (3.13), although the final expression may be rather involved.

Remark 3.8. Notice, that $\hat{\theta}_n$ need not be the maximum likelihood estimator (MLE) of θ; it is only proved that $\hat{\theta}_n$ is a maximum in a neighbourhood of θ_0. To prove that $\hat{\theta}_n$ is the MLE other (non-local) techniques must be used, e.g. as in Wald (1949) or Ivanov (1976). If the likelihood equation has a unique solution, then $\hat{\theta}_n$ must obviously coincide with the MLE.

Remark 3.9. Then inversion of a power series f, which is locally one-to-one may be obtained recursively by differentiation of f^{-1}, expressing the derivatives in terms of the derivatives of f. An explicit formula in the one-dimensional case is given in Skovgaard (1980). In the multivariate case Bolotov & Yuzhakov (1978) gives an explicit formula in terms of coordinates even for implicit functions, but no coordinate - free version seems to be known.

Corollary 3.10. Let the assumptions of Theorem 3.5 be fulfilled, and let $g \in C^p(\Theta,W)$ be a fixed function satisfying

$$D g(\theta_0) \quad \text{is non - singular} \tag{3.16}$$

If also $\| \Sigma_1^{-1} \| = o(1)$ then the distribution of $g(\hat{\theta}_n)$ may be expanded in an Edgeworth series of the form (3.10) replacing (3.13) by the stochastic expansion

$$D g(\theta_0) Y_1 + \sum_{j=2}^{p-1} \tilde{A}_j(Y_1,\ldots,Y_j) \tag{3.17}$$

where

$$\tilde{A}_j(Y_1,\ldots,Y_j)$$

$$= \sum_{\mu \in T(j)} D^{\Sigma \mu_i} g(\theta_0)[Y_1^{,\mu_1},\ldots,A_j(Y_1,\ldots,Y_j)^{,\mu_j}] / \prod_{i=1}^{j} \mu_i! \tag{3.18}$$

where $T(j) = \{(\mu_1,\ldots,\mu_j) \in \mathbb{N}_0^j \mid \Sigma i\mu_i = j\}$.

Proof of Theorem 3.5. The likelihood equation

$$D \log f(x;\theta) = 0 \tag{3.19}$$

may be expanded around $\theta = \theta_0$ yielding

$$S_1 + \sum_{j=2}^{p} (E_j + S_j)\,(\theta-\theta_0)^{\cdot j-1}/(j-1)! + R_1(\theta-\theta_0) = 0 \tag{3.20}$$

where $R_1(\theta-\theta_0)$ is stochastic. Write

$$Z = \Sigma_1^{\frac{1}{2}}\,(\theta-\theta_0),\ U_1 = \Sigma_1^{-\frac{1}{2}}\,S_1\ ,\ U_2 = B(S_1,\ldots,S_p)$$

where $B\colon B_1(V,\mathbb{R}) \times\ldots\times B_p(V,\mathbb{R}) \to V_2$ is a linear mapping into a Euclidean space V_2, and (U_1,U_2) is a normalization of $(S_1,\ldots,S_p)$, i.e. dim V + dim V_2 equals the dimension of the support of $(S_1,\ldots,S_p)$ and the variance of (U_1,U_2) is the identity on $V \times V_2$. Define

$$g\colon V \times V \times V_2 \to V$$

$$g(z,u_1,u_2) = u_1 + \sum_{j=2}^{p} \Sigma_1^{-\frac{1}{2}}(E_j + S_j(u_1,u_2))(\Sigma_1^{-\frac{1}{2}}(z)^{\cdot j-1})/(j-1)! + R_2(z) \tag{3.21}$$

where $R_2(z) = \Sigma_1^{-\frac{1}{2}}\,R_1(\Sigma_1^{-\frac{1}{2}}(z))$ and $(S_1,\ldots,S_p)(u_1,u_2)$ is the solutions of $(U_1,U_2)(S_1,\ldots,S_p) = (u_1,u_2)$ belonging to the affine support of $(S_1,\ldots,S_p)$. Thus (3.20) may be written

$$g(Z, U_1, U_2) = 0$$

Using Assumptions 3.2 (and 2.1) we obtain

$$g(0,0,0) = 0\ ,\ D\,g\,(0,0,0)(z,u_1,u_2) = u_1 - z$$

$$\| D^k g(0,0,0)(z,u_1,u_2)^{\cdot k} \|$$

$$\leq \| \Sigma_1^{-\frac{1}{2}}\,E_{k+1}(\Sigma_1^{-\frac{1}{2}}(z)^{\cdot k}) \| + k\,\| \Sigma_1^{-\frac{1}{2}}\,S_k(u_1,u_2)(\Sigma_1^{-\frac{1}{2}}(z)^{\cdot k-1}) \|$$

$$\leq \| E_{k+1} (\Sigma_1^{-\frac{1}{2}})',^{k+1} \| \| z \|^k + k \| S_k(u_1,u_2) \circ (\Sigma_1^{-\frac{1}{2}})',^k \| \| z \|^{k-1}$$

$$= O(\lambda_n^{k-1}) \text{ if } \| z \| \leq 1 , \| (u_1,u_2) \| \leq 1 , k \leq p \qquad (3.23)$$

Here we have used the fact that, since the variance of (U_1,U_2) is the identity, then the differential, D_k say, of $S_k(u_1,u_2) \circ (\Sigma_1^{-\frac{1}{2}})',^k$ with respect to (u_1,u_2) satisfies

$$\| D_k \|^2 = \| D_k \circ D_k^* \| = \| V\{S_k(U_1,U_2) \circ (\Sigma_1^{-\frac{1}{2}})',^k\} \|$$

$$= \| \Sigma_k \circ (\Sigma_1^{-\frac{1}{2}})',^{2k} \| \qquad (3.24)$$

Using (3.23) and (3.8) in (3.21) we obtain

$$g(z_1,u_1,u_2) - g(z_2,u_1,u_2) = - (z_1-z_2) + o(\sqrt{\lambda_n}) \| z_1 - z_2 \|$$

$$\text{if } z_1,z_2 \in H_n(\alpha) , \| (u_1,u_2) \| \leq \rho_n(\alpha) , X \in D_n \qquad (3.25)$$

because $\rho_n^m(\alpha)\lambda_n = o(1)$ for any $\alpha > 0$, $m > 0$. Thus for any fixed (u_1,u_2) $(\|(u_1,u_2)\| \leq \rho_n(\alpha))$ and n sufficiently large there is with probability $1 - o(\beta_n)$ at most one solution $z \in H_n(\alpha_1)$ to the likelihood equation, $\alpha_1 < \alpha$.

Let $\delta > 0$ be fixed. Then if $\| (u_1,u_2) \| \leq \rho_n(\alpha_1)$, $\alpha_1 < \alpha$, $\| z - u_1 \| < \delta$ and n is sufficiently large we have $z \in H_n(\alpha)$. To prove the existence of a solution $z \in H_n(\alpha)$ of (3.22) we apply Brauer's fixpoint theorem to the function

$$\tilde{g}(y) = g(u_1+y,u_1,u_2) + y , \| y \| < \varepsilon , 0 < \varepsilon < \delta.$$

By the remark above and (3.25) we have

$$\| y \| < \varepsilon \Rightarrow \| \tilde{g}(y) \| < \varepsilon.$$

because $g(u_1,u_1,u_2) = u_1 + o(1)$. Also, by (3.25)

$$\| \tilde{g}(y_1) - \tilde{g}(y_2) \| = o(1) \| y_1 - y_2 \| , \| y_1 \| , \| y_2 \| < \varepsilon$$

proving, that $\tilde{g}$ has a fixpoint in $\{\| y \| < \varepsilon\}$, implying the existence of a solution $z(= y + u_1) \in H_n(\alpha)$ to the likelihood equation when $X \in D_n$ and n is sufficiently large.

By the uniqueness of power series expansions, (3.13) must be the p-1 order Taylor series expansion of $\hat{\theta}_n = \Sigma_1^{-\frac{1}{2}} Z_n$ in terms of $(S_1,\ldots,S_{p-1})$ around zero, where Z_n is the solution of (3.22), and hence $\hat{\theta}_n$ a solution of the likelihood equation (3.19). Thus, it only remains to be proved, that the derivatives of Z_n with respect to (U_1,U_2) satisfies Assumptions 3.1 of Skovgaard (1980), since the expansion (4.5) of Skovgaard (1980) then implies that $\hat{\theta}_n$ locally maximizes the likelihood function.

Write $u = (u_1,u_2)$ and let $z = \psi(u)$ be the solution (in $H_n(\alpha)$) of the equation (3.22). Also, if $\omega(u) = (\psi(u),u)$

$$D^k(g\circ\omega)(u_0) = 0 , k \geq 0, \| u_0 \| \leq \rho_n(\alpha_1) \qquad (3.26)$$

and using a general formula (see Federer (1969), 3.1.11)

$$D^k(g\circ\omega)(u_0)(u'^k)$$

$$= \sum_{\mu\in T(k)} k! D^{\Sigma\mu_i} g(z_0,u_0)[D\omega(u_o)(u)'^{\mu_1},\ldots,D^k\omega(u_o)(u'^k)'^{\mu_k}] / \prod_{i=1}^{k} \mu_i!(i!)^{\mu_i}$$

$$, z_0 = \psi(u_0) \qquad (3.27)$$

Where T(k) is given in Corollary 3.10. From (3.27) and (3.23) we obtain by induction

$$D\,\psi(0)(u) = u_1$$

$$\|D^k\,\psi(0)\| \le \sum_{\mu\in T'(k)} k!\,\|D^{\Sigma\mu_i}\,g(0,0)\| \prod_{i=1}^{k} (\|D^i\omega(0)\|/i!)^{\mu_i}/\mu_i!$$

$$= \sum_{\mu\in T'(k)} O(\lambda_n^{\Sigma\mu_i - 1}) \prod_{i=1}^{k} O(\lambda_n^{i-1})$$

$$= O(\lambda_n^{k-1}) \quad , \quad 2 \le k \le p-1 \tag{3.28}$$

where $T'(k) = T(k)\setminus\{(0,\ldots,0,1)\}$.

Using (3.8) and (3.21) it is seen that, if $\|u\|, \|z\| < \rho_n(\alpha)$, then $\|D^k g(z,u)\| = O(\lambda_n^{k-1})$, $k \le p-1$ and $\|D^p g(z,u)\| = o(\beta_n)$ if $X \in D_n$, and as above it follows, that

$$\|D^p\psi(u)\| = o(\beta_n) \text{ uniformly in } \{\|u\| \le \rho_n(\alpha)\} \tag{3.29}$$

By (3.28) and (3.29), Assumptions 3.1 in Skovgaard (1980), and hence our Theorem 3.5,is proved. □

Proof of Corollary 3.10.

The formula (3.18) is easily obtained using the formula for derivatives of composite functions, see Federer (1969), 3.1.11. By this formula and the assumption $\|\Sigma_1^{-1}\| = o(1)$, which says that the eigenvalues of the Fisher-information tends uniformly to infinity, it follows, that the assumptions of Theorem 3.2 in Skovgaard (1980) are fulfilled, proving the corollary. □

4. Exponential family models

In this section we consider (for each $n \in \mathbb{N}$) a setup of the form given below. Assume that E is a finite dimensional Euclidean space, $\mathcal{E}$ the Borel σ-field on E and μ a measure on $(E,\mathcal{E})$. Define

$$\psi(\eta) = \log \int \exp\{<\eta,x>\} d\mu(x), \quad \eta \in H \subseteq E \tag{4.1}$$

where H is the subset of E for which the integral is positive and finite. Define the family (P_η), $\eta \in H$ of probability measures on $(E,\mathcal{E})$ by

$$(dP_\eta/d\mu)(x) = f(x;\eta) = \exp\{<\eta,x>-\psi(\eta)\} \tag{4.2}$$

The model we shall consider is given by a differentiabel parametrization

$$\beta \in C^p(\Theta,H) \ , \ \eta = \beta(\theta) \ , \ \theta \in \Theta \subseteq V \tag{4.3}$$

where V is a finite dimensional Euclidean space independent of n, and usually of lower dimension than E. The cumulants of P_{η_0}, $\eta_0 = \beta(\theta_0)$, are

$$\chi_k = D^k\psi(\eta_0) \ , \ k \in \mathbb{N} \tag{4.4}$$

Also

$$D \log L(\eta(\theta_0)) = < x - D\psi(\eta_0) \ , \ D\beta(\theta_0) > \tag{4.5}$$

and accordingly

$$-E_k = \sum_{\nu \in T'(k)} k! \, \chi_{\Sigma\nu_i} \circ [(D\beta_0)^{\nu_1},\ldots,(D^k\beta_0)^{\nu_k}] / \prod_{i=1}^{k} \nu_i!(i!)^{\nu_i} \tag{4.6}$$

$$S_k = < x - \chi_1, D^k\beta_0 > \tag{4.7}$$

$$\Sigma_k = \chi_2 \circ (D^k\beta_0, D^k\beta_0) \tag{4.8}$$

where E_k, S_k amd Σ_k are defined in (2.13), (2.14) and (2.15), $D^k\beta_0 = D^k\beta(\theta_0)$ and $T'(k) = T(k) \setminus \{(0,\ldots,0,1)\}$.

Thus the approximate cumulants in Theorem 3.5 may be expressed explicitely in terms of (χ_k), $(D^k\beta_0)$, $k \geq 1$. Some of these cumulants are given in Section 6 in a coordinate version. The expressions may be somewhat simplified using a coordinate-free notation, but for computations this is not useful. Recall, that for fixed p in (3.13), only the first p cumulants are needed.

Remark 4.1. There are a number of situations, where the expression (3.13) and its cumulants are considerably simpler. These include

(a) A canonical model, i.e. β is affine. Then

$$-E_k = \chi_k \circ (D\beta_0, {}^k) \quad , \quad k \geq 2$$

$$S_k = 0 \quad , \quad \Sigma_k = 0 \quad , \quad k \geq 2 \tag{4.9}$$

(b) An affine mean value structure, i.e. $(D\psi)\circ\beta$ is affine Then

$$-E_k = k\ \chi_2 \circ (D\beta_0, D^{k-1}\beta_0) \quad , \quad k \geq 3 \tag{4.10}$$

where E_k is understood to be symmetric.

(c) The normal case (with fixed variance), where

$$\chi_k = 0 \quad , \quad k \geq 3$$

The normal regression models will be discussed further in the next section.

If, in particular, both (a) and (b) are fulfilled, then the MLE is an affine function of the minimal sufficient statistic S_1, and the transformation of an Edgeworth expansion of S_1 to an Edgeworth expansion of the MLE is trivially valid.

5. Normal non-linear regression.

Consider a sequence $X_1, X_2, \ldots$ of independent random vectors, X_i normally distributed on $\mathbb{R}^m$ with mean $\mu_i(\theta) \in \mathbb{R}^m$ and variance $\Sigma = \sigma^2 \Sigma_0$, $\sigma^2 > 0$, $\Sigma_0 \in B_2(\mathbb{R}^m, \mathbb{R})$. Σ_0 is supposed to be known, $\theta \in V$ unknown. Whether σ^2 is known or unknown is immaterial, when considering maximum likelihood estimation of θ. We shall consider σ^2 as known for simplicity. With notation as in the previous sections, we have

$$\log f(x;\theta) = \text{const} - \tfrac{1}{2} \sum_{i=1}^{n} \Sigma^{-1}(x_i - \mu_i(\theta), x_i - \mu_i(\theta)) \qquad (5.1)$$

from which we derive

$$-E_k = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k-1} \binom{k}{j} \Sigma^{-1} \circ (D^j \mu_i(\theta_0), D^j \mu_i(\theta_0)), \quad k \geq 2 \qquad (5.2)$$

$$S_k = \sum_{i=1}^{n} \Sigma^{-1}(X_i - \mu_i(\theta_0), D^k \mu_i(\theta_0)) \quad k \geq 1 \qquad (5.3)$$

$$\Sigma_k = \sum_{i=1}^{n} \Sigma^{-1} \circ (D^k \mu_i(\theta_0), D^k \mu_i(\theta_0)), \quad k \geq 1 \qquad (5.4)$$

Since this class of models is widely used, we shall in somewhat more details investigate under which conditions Assumptions 3.2 are fulfilled. Notice, that Assumptions 3.1 are fulfilled with

$\beta_n = 0$, because $(S_1,\ldots,S_p)$ are exactly normally distributed. Of Assumptions 2.1 only IV needs to be checked.

Lemma 5.1. Let (E_k) and (S_k) be given by (5.2) and (5.3). Then (3.5) and (3.7) implies (3.6).

Proof. We shall prove, that

$$| E_{k+1}(v'^{k+1}) | = (\Sigma_1(v,v))^{(k+1)/2}\, O(\lambda_n^{k-1}),\ 2 \le k \le p-1,\ v \in V \quad (5.5)$$

By (5.2) and Cauchy-Schwarz inequality we have

$$| E_{k+1}(v'^{k+1}) |$$

$$= \frac{1}{2} \sum_{j=1}^{k} \binom{k+1}{j} \sum_{i=1}^{n} \Sigma^{-1}(D^j \mu_i(\theta_0)(v'^j),\ D^{k+1-j}\mu_i(\theta_0)(v'^{k+1-j}))$$

$$\le \frac{1}{2} \sum_{j=1}^{k} \binom{k+1}{j} [\sum_{i=1}^{n} \Sigma_j(v'^{2j})]^{\frac{1}{2}} [\sum_{i=1} \Sigma_{k+1-j}(v'^{2k+2-2j})]^{\frac{1}{2}}$$

$$= \frac{1}{2} \sum_{j=1}^{k} \binom{k+1}{j} \Sigma_1(v,v)^{(k+1)/2}\, O(\lambda_n^{j-1})\, O(\lambda_n^{k-j})$$

$$= (\Sigma_1(v,v))^{(k+1)/2}\, O(\lambda_n^{k-1}) \qquad \square$$

Notice, that since $(S_1,\ldots,S_p)$ is exactly normally distributed, the sequence (λ_n) may be chosen as any sequence, which is $o(1)$. Next, we shall prove that also (3.8) may be deduced under simple conditions.

Lemma 5.2. Suppose, that the functions (μ_i) are analytic in a neighbourhood of θ_0, and that (3.7) holds uniformly in $j \ge 2$, then Assumptions 3.2 hold with $\beta_n = \lambda_n^{p-2}$

Proof. For sufficiently large n, $\log f(x;\theta)$ will coincide with its Taylor series expansion around $\theta = \theta_0$, when $\|\theta - \theta_0\|$ is less than the radius of convergence. Hence

$$D^p \log f(x;\theta) \circ (\Sigma_1^{-\frac{1}{2}})',^p$$

$$= D^p \log f(x;\theta_0) \circ (\Sigma_1^{-\frac{1}{2}})',^p + \sum_{j=p+1}^{\infty} D^j \log f(x;\theta_0)$$

$$((\Sigma_1^{-\frac{1}{2}})',^p, (\theta-\theta_0)',^{j-p})/(j-p)! \qquad (5.6)$$

on the set

$$M_c = \{\theta \in V \mid \|\Sigma_1^{\frac{1}{2}}(\theta-\theta_0)\|^{j-p} \|D^j \log f(x;\theta_0) \circ (\Sigma^{-\frac{1}{2}})',^j\|/(j-p)! < c^{j-p}\} \qquad (5.7)$$

for any $c \in]0,1[$. Rewriting (5.6) we obtain

$$D^p \log f(x;\theta) \circ (\Sigma_1^{-\frac{1}{2}})',^p$$

$$= (E_p+S_p) \circ (\Sigma^{-\frac{1}{2}})',^p + \sum_{j=p+1}^{\infty} (E_j \circ (\Sigma^{-\frac{1}{2}})',^j)(z',^{j-p})/(j-p)!$$

$$+ \sum_{j=p+1}^{\infty} (S_j \circ (\Sigma_1^{-\frac{1}{2}})',^j)(z',^{j-p})/(j-p)! \ , \ \theta \in M_c \qquad (5.8)$$

where $z = \Sigma_1^{\frac{1}{2}} (\theta-\theta_0)$.

By a slight modification of the proof of Lemma 5.1 it follows, that (3.6) holds uniformly in $j \geq 2$, hence the first sum in (5.7) is $O(\lambda_n^{p-1})$ if $z \in H_n(\alpha)$.

The next step is to obtain bounds on (S_j), $j \geq p$ holding with probability $1 - o(\beta_n) = 1 - o(\lambda_n^{p-2})$. Let d be the dimension of V. Then $D^j \underset{\sim}{\mu}(\theta_0) \in B_j(V, \mathbb{R}^{mn})$, $\underset{\sim}{\mu}(\theta) = (\mu_1(\theta),\ldots,\mu_n(\theta))$, spans a d^j-dimensional subspace, L_j say, of $\mathbb{R}^{mn}$. Let p_j denote the

projection on L_j w.r.t. the metric Δ on $\mathbb{R}^{mn}$ induced by the metric Σ^{-1} on each component $\mathbb{R}^m$. Then, using (5.3),

$$S_j = \Delta\,(p_j(\underset{\sim}{X} - \underset{\sim}{\mu}(\theta_0)), D^j \underset{\sim}{\mu}(\theta_0)),\ \underset{\sim}{X} = (X_1,\ldots,X_n) \tag{5.8}$$

and

$$\begin{aligned} \| S_j \circ (\Sigma_1^{-\frac{1}{2}})',^j \| &= \| \Delta(p_j(\underset{\sim}{X}-\underset{\sim}{\mu}(\theta_0)), D^j\underset{\sim}{\mu}(\theta_0) \circ (\Sigma_1^{-\frac{1}{2}})',^j) \| \\ &\le \Delta((p_j(\underset{\sim}{X}-\underset{\sim}{\mu}(\theta_0)',^2)^{\frac{1}{2}} \| \Delta(D^j\underset{\sim}{\mu}(\theta_0)\circ(\Sigma_1^{-\frac{1}{2}})',^j)',^2) \|^{\frac{1}{2}} \\ &= \Delta((p_j(\underset{\sim}{X}-\underset{\sim}{\mu}(\theta_0)))',^2)^{\frac{1}{2}} \| \Sigma_j \circ (\Sigma_1^{-\frac{1}{2}})',^{2j} \| \end{aligned} \tag{5.9}$$

The first factor is the Δ-norm of a d^j-dimensional normally distributed random vector with mean zero and variance Δ^{-1} restricted to L_j. Thus by Lemma 4.1 in Skovgaard (1980) we have for any $K_j > 0$

$$\begin{aligned} &P\{\Delta((p_j(\underset{\sim}{X}-\underset{\sim}{\mu}(\theta_0)))',^2)^{\frac{1}{2}} > K_j\} \\ &\le e^{-\frac{1}{2}K_j^2} (K_j^{d^j-2}/\Gamma(d^j/2) + \sqrt{2}^{d^j-2}) \end{aligned} \tag{5.10}$$

Chosing $K_j = K \sqrt{\lambda}_n^{-(j-p)}$, $K > 0$, we obtain

$$\begin{aligned} &P \bigcup_{j=p+1}^{\infty} \{\Delta((p_j(\underset{\sim}{X}-\underset{\sim}{\mu}(\theta_0)))',^2) > K_j\} \\ &\le \sum_{j=p+1}^{\infty} \exp\{-\tfrac{1}{2}K^2 \sqrt{\lambda}_n^{p-j}\}(K_j^{d^j-2}/\Gamma(d^j/2) + \sqrt{2}^{d^j}) \end{aligned} \tag{5.11}$$

which decreases towards zero at exponential rate in $\sqrt{\lambda}_n^{-1}$.

Combining this with (5.9), we have

$$\| S_j \circ (\Sigma_1^{-\frac{1}{2}})', j \| = K \sqrt{\lambda}_n^{p-j} \, O(\lambda_n^{j-1})$$

with probability $1 - o(\lambda_n^{p-2})$, implying that the second sum in (5.8) is $O(\lambda_n^{p-\frac{1}{2}}) = o(\beta_n)$ with probability $1 - o(\beta_n)$, $\beta_n = \lambda_n^{p-2}$, on the set $\theta_0 + \Sigma_1^{-\frac{1}{2}} H_n(\alpha)$. □

Remark 5.3. The conditions of Lemma 5.2 may be stated in the following form . The functions (μ_i) are analytic, and (by Remark 3.4)

$$\Big| \sum_{i=1}^{n} \Sigma^{-1} (D^j \mu_i(\theta_0)(v', j), 2 \Big|^{\frac{1}{2}} / \Big(\sum_{i=1}^{n} \Sigma^{-1} (D\mu_i(\theta_0)(v), 2) \Big)^{j/2}$$

$$= O(\lambda_n^{j-1}) \text{ uniformly in } v \in V \text{ and } j \geq 2 \qquad (5.12)$$

Remark 5.4. Another interesting case, closely connected with the one discussed above, occurs if, in the non-linear regression models discribed above, we fix n, and consider the limiting behaviour as $\sigma^2 \to 0$. It is quit trivial to check, that Assumptions 3.2 are fulfilled with $\lambda = \sigma$, and hence that the conclusion of Theorem 3.5 holds. This proves that the asymptotic results may be applied if the variance is small, even if the number of observations is small.

An example: exponential regression.

Let $X_1, \ldots, X_n$ be independent, $X_i \in \mathbb{R}$ normally distributed with mean

$$\mu_i(\theta_1, \theta_2) = \theta_1 e^{\theta_2 t_i} , \quad \theta_1 > 0 , \theta_2 \in \mathbb{R} , t_i \in \mathbb{R} \qquad (5.13)$$

and variance $\sigma^2 > 0$. The conditions of Lemma 5.2 are verified as

follows. First note that the functions (μ_i) and hence the likelihood functions are analytic. Let $\eta = (\eta_1,\eta_2) \in \mathbb{R}^2$, $\|\eta\| = 1$. Then for any $\theta = (\theta_1,\theta_2)$

$$[\sum_{i=1}^{n}(D^j\mu_i(\theta)(\eta'^j))^2 / \sigma^2]/[\sum_{i=1}^{n}(D\mu_i(\theta)(\eta))^2/\sigma^2]^j$$

$$= \sum_{i=1}^{n}(\eta_1\eta_2^{j-1} e^{\theta_2 t_i} + \eta_2^j \theta_1 t_i^j e^{\theta_2 t_i})^2$$

$$/([\sum_{i=1}^{n}(\eta_1 e^{\theta_2 t_i} + \eta_2\theta_1 t_i e^{\theta_2 t_i})^2]^j(\sigma^2)^{j-1})$$

$$\leq (\max\{t_i \mid i = 1,\ldots,n\})^{2(j-1)} / \bar{\Sigma}_1(\eta,\eta)^{j-1} \qquad (5.14)$$

where $\Sigma_1(\eta,\eta) = \sum_{i=1}^{n}(D\mu_i(\theta)(\eta))^2/\sigma^2$ is the Fisher-information of η. Thus if

$$\lambda_n = (\max\{t_i \mid i=1,\ldots,n\})\,\|\Sigma_1^{-\frac{1}{2}}\| = o(1)$$

then Theorem 3.5 is applicable. E.g. if $\theta_2 > 0$ and $t_i = i$, then λ_n will decrease exponentially fast. Thus, if one has observations at equidistant points (t's) a very good agreement between the correct distribution and the approximations may be expected with relatively few points, but, of course, this can only be proved by estimating the difference.

Although this example is of practical interest in itself, it is unusually simple. The condition (5.12) is however so simple, that further simplification of importance is hardly obtainable.

6. A coordinate version of the results.

We shall use a notation commonly used in tensor calculus. An array $(M_{i_1 \ldots i_m})$, $i_1,\ldots,i_m \in \{1,\ldots,k\}$ belonging to $(\mathbb{R}^k)^m$ will be written $M_{i_1 \ldots i_m}$ without explicitely stating the range of the indices. Some indices will be written as superscripts some as subscripts. These corresponds to contravariant and covariant tensors, but the destinction is not important for this application. We also use the standard summation convention, i.e. if an index appears twice in a term, summation w.r.t. this index over its range is understood.

First we shall give the first four terms of the expansion (3.13). Define

$$E^{i_1 \ldots i_m} = E\{\frac{d}{d\theta_{i_1}} \ldots \frac{d}{d\theta_{i_m}} \log f(X;\theta) \Big|_{\theta=\theta_0}\} \tag{7.1}$$

$$S^{i_1 \ldots i_m} = (\frac{d}{d\theta_{i_1}} \ldots \frac{d}{d\theta_{i_1}} \log f(X;\theta) \Big|_{\theta=\theta_0}) - E^{i_1 \ldots i_m} \tag{7.2}$$

where $\theta = (\theta_1,\ldots,\theta_k) \in \mathbb{R}^k$. Also

$$Y_i^{i_2 \ldots i_m} = g_{ij} S^{ji_2 \ldots i_m} \tag{7.3}$$

$$F_i^{i_2 \ldots i_m} = -g_{ij} E^{ji_2 \ldots i_m} \tag{7.4}$$

where (g_{ij}) is the inverse of $(g^{ij}) = (-E^{ij})$, i.e. the inverse Fisher-information.

The first four terms of (3.13)

$$(A_1)_i = Y_i$$

$$(A_2)_i = Y_i^j Y_j - \frac{1}{2} F_i^{jk} Y_j Y_k$$

$$(A_3)_i = Y_i^j Y_j^k Y_k - \frac{1}{2} Y_i^j F_j^{kl} Y_k Y_l - F_i^{jk} Y_j Y_k^l Y_l + \frac{1}{2} Y_i^{jk} Y_j Y_k$$
$$- \frac{1}{6} F_i^{jkl} Y_j Y_k Y_l + \frac{1}{2} F_i^{jk} Y_j F_k^{lm} Y_l Y_m$$

$$(A_4)_i = Y_i^j Y_j^k Y_k^l Y_l - \frac{1}{2} Y_i^j Y_j^k F_k^{lm} Y_l Y_m - Y_i^j F_j^{kl} Y_k Y_l^m Y_m - \frac{1}{2} F_i^{jk} Y_j^l Y_l Y_k^m Y_m$$
$$- F_i^{jk} Y_j Y_k^l Y_l^m Y_m + \frac{1}{2} Y_i^j Y_j^{kl} Y_k Y_l + Y_i^{jk} Y_j Y_k^l Y_l - \frac{1}{6} Y_i^j F_j^{klm} Y_k Y_l Y_m$$
$$- \frac{1}{2} F_i^{jkl} Y_j Y_k Y_l^m Y_m + \frac{1}{6} Y_i^{jkl} Y_j Y_k Y_l + \frac{1}{2} Y_i^j F_j^{kl} Y_k F_l^{mn} Y_m Y_n$$
$$+ \frac{1}{2} F_i^{jk} Y_j^l Y_l F_k^{mn} Y_m Y_n + F_i^{jk} Y_j F_k^{lm} Y_l Y_m^n Y_n + \frac{1}{2} F_i^{jk} Y_j Y_k^l F_l^{mn} Y_m Y_n$$
$$- \frac{1}{2} Y_i^{jk} Y_j F_k^{lm} Y_l Y_m - \frac{1}{2} F_i^{jk} Y_j Y_k^{lm} Y_l Y_m - \frac{1}{24} F_i^{jklm} Y_j Y_k Y_l Y_m$$
$$+ \frac{1}{6} F_i^{jk} Y_j F_k^{lmn} Y_l Y_m Y_n + \frac{1}{4} F_i^{jkl} Y_j Y_k F_l^{mn} Y_m Y_n$$
$$- \frac{1}{3} F_i^{jk} Y_j F_k^{lm} Y_l F_m^{no} Y_n Y_o - \frac{1}{8} F_i^{jk} F_j^{lm} Y_l Y_m F_k^{no} Y_n Y_o \qquad (7.5)$$

Approximate cumulants in the exponential family models.

Using the method described in Remark 3.7 it is straight forward to calculate the approximate cumulants (κ_j) of (7.5). Let $(\kappa_m)_{i_1 \ldots i_j}$ denote the j'th cumulant of the m'th approximation, i.e. with q = m + 1 in (3.11) Thus $(\kappa_1)_i$ and $(\kappa_1)_{ij}$ denote mean and variance in the first (normal) approximation. With obvious modification of the notation in section 4 we define

$$[i,j] = (\chi_2)^{\alpha\beta} (D\beta_0)_\alpha^i (D\beta_0)_\beta^j = g^{ij}$$

$$[i,j,k] = \chi_3^{\alpha\beta\gamma} (D\beta_0)_\alpha^i (D\beta_0)_\beta^j (D\beta_0)_\gamma^k$$

$$[i,jk] = \chi_2^{\alpha\beta} (D\beta_0)_\alpha^i (D^2\beta_0)_\beta^{jk} \qquad \text{etc.}$$

where $(D_\alpha^m \beta_0)^{i_1 \cdots i_m} = \frac{d}{d\theta_{i_1}} \cdots \frac{d}{d\theta_{i_m}} \beta(\theta)_\alpha \Big|_{\theta=\theta_0}$. Then we have

m = 1:

$$(\kappa_1)_i = 0 \ , \ (\kappa_1)_{ij} = g_{ij} = [i,j]$$

m = 2:

$$(\kappa_2)_i = -\frac{1}{2} g_{ij}([j,kl] + [j,k,l])g_{kl}$$

$$(\kappa_2)_{ij} = (\kappa_1)_{ij} = g_{ij}$$

$$(\kappa_2)_{ijk} = -\operatorname{sym}\{g_{il} g_{jm} g_{kn}(2[l,m,n] + 3[l,mn])\}$$

m = 3:

$$(\kappa_3)_i = (\kappa_2)_i$$

$$\begin{aligned}(\kappa_3)_{ij} &= g_{ij} + g_{ik} g_{jl} ([km,nl] - [kl,mn] - [k,m,n,l])g_{mn} \\ &+ \operatorname{sym}\{g_{ik} g_{jl} g_{mn}(- [k,lmn] - [k,l,mn] - [m,n,kl] \\ &- 2[k,m,nl]) + g_{ik} g_{jl} g_{mn} g_{op} ([k,l,m][n,o,p] \\ &+ \tfrac{3}{2} [k,m,o][l,n,p] \\ &+ [k,l,m][n,op] + [k,lm][n,o,p] + [m,kl][n,o,p] \\ &+ 2[k,m,o][l,np] + 2[k,m,o][n,lp] + 2[k,mo][n,lp] \\ &- [m,ko][n,lp] + \tfrac{1}{2} [k,mo][l,np] + [m,kl][n,op] \\ &+ [k,lm][n,op])\}\end{aligned}$$

$$(\kappa_3)_{ijk} = (\kappa_2)_{ijk}$$

$$(\kappa_4)_{ijkl} = \mathrm{sym}\{g_{i\alpha}\, g_{j\beta}\, g_{k\gamma}\, g_{l\,\delta}(-3\,[\alpha,\beta,\gamma,\delta] - 4\,[\alpha,\beta\gamma\delta]$$

$$- 12[\alpha,\beta,\gamma\delta] + 12[\alpha,\beta,m]g_{mn}[n,\gamma,\delta] + 12[\alpha,\beta,m]g_{mn}[n,\gamma\delta]$$

$$+ 24[\alpha,\beta,m]g_{mn}[\gamma,\delta n] + 12[\alpha,\beta m]g_{mn}[n,\gamma\delta]$$

$$+ 12[\alpha,\beta m]g_{mn}[\gamma,\delta n])\}. \qquad (7.6)$$

where sym{...} means the avarage over all permutations of the indices appearing on the left hand side on the equation. Actually taking this avarage is not necessary in applications, because the appearance of the cumulants in (3.11) is symmetric in their indices. This fact is a considerable relief in calculations.

The variance term for $m = 3$ may be identified with that given in Efron (1975) in the one-dimensional case and with its multivariate generalization in L.T.Skovgaard (1979). That our formula seems more complicated is only because of the less directly computable terms appearing in the above mentioned papers. All the terms, except κ_4, may be found in Shenton & Bowman (1977). Notice, however that their square brackets have a meaning different from ours.

An interesting feature of the correction terms for $m = 2$ (i.e. the first correction to the normal distribution) is, that since $\gamma_{il}\,\gamma_{jm}\,\gamma_{kn}\,[l,m,n]$ is invariant under reparametrizations in the one-dimensional case, and in the multivariate case its range is invariant, then the first correction term of (3.11) cannot be removed by a reparametrization, unless this invariant vanishes, e.g. if the third cumulant of the exponential family is zero.

In the case of a non-linear normal regression model (Section 5) the cumulants (7.6) are still valid, but important simplification is achieved, because only the square bracket factors of the form $[i_1 \ldots i_2, j_1 \ldots j_\beta]$ are different from zero. Thus [i,j,k],[i,j,kl], etc. vanish.

Acknowledgement.

I wish to thank Steffen L. Lauritzen for useful comments and discussions on the subject.

References.

Bhattacharya, R.N. & Ghosh, J.K. (1978). On the validity of the formal Edgeworth expansion. Ann. Statist. 6, 434-451.

Bolotov, V.A. & Yuzhakov, A.P. (1978). A generalization of the inversion formulas of systems of power series in systems of implicit functions. (Russian). Mat. Zametki 23, 47-54.

Brillinger, D.R. (1975). Time series: data analysis and theory. Holt, Rinehart and Winston, New York.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). Ann. Statist. 3, 1189-1242.

Federer, H. (1969). Geometric Measure Theory. Springer, New York.

Ivanov, A.V. (1976). An asymptotic expansion for the distribution of the least squares estimator of the non-linear regression parameter. Theor. Probability Appl. 21, 557-570.

Leonov, V.P. & Shiryaev, A.N. (1959). On a method of calculation of semi-invariants. Theor. Probability Appl. 4, 319-329.

Shenton, L.R. & Bowman, K.O. (1977). Maximum likelihood estimation in small samples. Griffin, London,

Skovgaard, L.T. (1979). The geometry of statistical models. A generalization of curvature. Research report 79/1, Statistical Research Unit, Copenhagen.

Skovgaard, I.M. (1980). Transformation of an Edgeworth expansion by a sequence of smooth functions. Preprint 1980, 2, Inst. Math. Stat., Univ. of Copenhagen.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. Ann. Math. Stat. 20, 595-601.

PREPRINTS 1979

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

No. 1 Edwards, David: Large Sample Tests for Stationarity and Reversibility in Finite Markov Chains.

No. 2 Andersson, Steen A. : Distribution of Maximal Invariants Using Proper Action and Quotient Measures.

No. 3 Johansen, Søren: A Note on the Welch-James Approximation to the Distribution of the Residual Sum of Squares in a Weighted Linear Regression.

No. 4 Björnsson, Ottó: Four Simple Characterizations of Standard Borel Spaces.

No. 5 Johansen, Søren: Some Comments on Robustness

No. 6 Hald, Anders: T.N. Thiele's Contributions to Statistics.

No. 7 Jacobsen, Martin: Markov Chains: Birth and Death Times with Conditional Independence.

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.