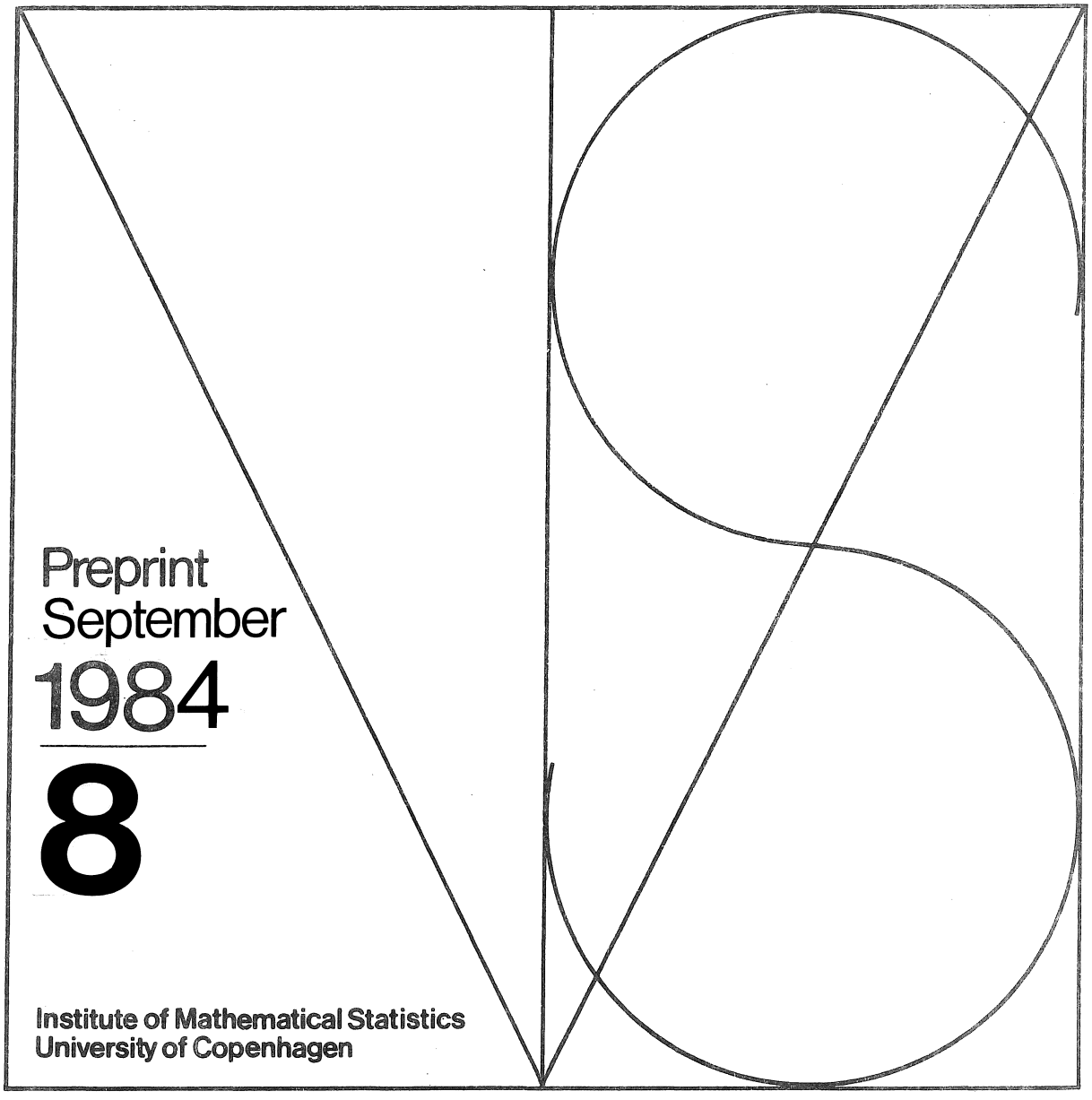
Philip Hougaard

Saddlepoint Approximations for Curved Exponential Families



Philip Hougaard*

SADDLEPOINT APPROXIMATIONS FOR CURVED EXPONENTIAL FAMILIES

Preprint 1984 No. 8

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

September 1984

ABSTRACT

An approximation to the density of the maximum likelihood estimator in curved exponential families is derived using a saddlepoint expansion. The approximation is particularly simple in nonlinear regression. An example is considered.

Key words: curved exponential family; nonlinear regression; saddlepoint approximation.

1. INTRODUCTION

Using a saddlepoint technique Field (1982) derived a small sample approximation for multivariate M-estimates. This approximation is, however, fairly complicated. It involves a so-called conjugate distribution determined by means of an integral equation. Also it requires the calculation of several mean values under this conjugate distribution. The aim of this note is to specialize the approximation to maximum likelihood estimators in curved exponential families, where the approximation is much simpler. The conjugate distribution is a member of the exponential family in which the curved exponential family is embedded and this distribution can easier be determined, because the integrals involved can be explicitly calculated and expressed by means of the normalizing constant. In nonlinear regression the approximation is even simpler, because it can be expressed directly, that is without involving the conjugate distribution. The approximation is the same as the one derived by Skovgaard (1981). For papers on saddlepoint approximations, see Barndorff-Nielsen (1980) and Daniels (1983).

The results are described in Section 2. In Section 3 we discuss the necessary regularity conditions for the results and in Section 4 the results are proved. An example is considered in Section 5.

2. RESULTS

Consider a curved exponential family, that is a p-dimensional curved submodel of a k-dimensional exponential family. The exponential family has density $f(x;\theta) = \exp\{\theta' t(x)\}\mu(dx)/\phi(\theta)$, where $\theta \in D \subseteq \mathbb{R}^k$, $\phi(\theta)$ the normalizing constant, μ a measure dominated by Lebesgue measure, $t(x)$ the k-dimensional sufficient statistic and the prime denotes transposing. The curved submodel is constructed by assuming that θ is a function $\theta(\beta)$ of a parameter $\beta \in B \subseteq \mathbb{R}^p$. For our purpose we need to know the value of $\phi(\theta)$ or equivalently $\chi(\theta) = \ell n\, \phi(\theta)$ for the full k-dimensional family.

For the asymptotics we consider n independent identically distributed variables $X_1,\ldots,X_n$ following the distribution above and then we let $n\to\infty$. In that case the sum $\Sigma_1^n t(X_i)$ or equivalently the average $\bar{t} = \Sigma_1^n t(X_i)/n$ is sufficient and follows an exponential family distribution. For the average the canonical parameter is $\theta_n = n\theta$ and the normalizing constant is $\phi_n(\theta_n) = \phi(\theta)^n = \phi(\theta_n/n)^n$.

The maximum likelihood estimator $\hat{\beta}$ is usually found by means of the likelihood equation, derived by differentiation of the logarithm of the likelihood function, giving

$$n\{\bar{t} - \tau(\theta(\beta))\}' d\theta/d\beta = 0,$$

where $\tau(\theta) = d\chi/d\theta = E_\theta t(X)$. This equation is a special case of Field (1982, eq. (1)), which says $\Sigma_i\, \psi_j(X_i,\beta) = 0$, $j = 1,..,p$. The true value is denoted β_0 and we let $\theta_0 = \theta(\beta_0)$. The value of θ corresponding to the conjugate distribution, to be defined below, will be denoted θ^*. By $d^2\chi/d\theta^{*2}$ we mean $d^2\chi/d\theta^2$ evaluated at $\theta = \theta^*$. This notation is fairly different from the notation of Field (1982).

We can then specialize the results of Theorem 1 in Field (1982) to maximum

likelihood estimation in curved exponential families.

Theorem 1 The curved exponential family.

Under regularity conditions, cf. Section 3, the density of $\hat{\beta}$ at β can, if $\theta(\beta)$ is an interior point in D, be expanded as

$$p_n(\beta) = (\frac{n}{2\pi})^{p/2} \exp[n\{\chi(\theta^*) - \chi(\theta_0) - \tau(\theta)'\frac{d\theta}{d\beta}\alpha\}]$$

$$[|\frac{d\theta'}{d\beta} \frac{d^2\chi}{d\theta^2} \frac{d\theta}{d\beta} - \{\tau(\theta^*) - \tau(\theta)\}'\frac{d^2\theta}{d\beta^2}|$$

$$|\frac{d\theta'}{d\beta} \frac{d^2\chi}{d\theta^{*2}} \frac{d\theta}{d\beta}|^{-\frac{1}{2}} + o(1/n)] ,$$

where $\theta^* = \theta_0 + d\theta/d\beta\ \alpha$ with α given by

$$\{\tau(\theta^*) - \tau(\theta)\}' \frac{d\theta}{d\beta} = 0.$$

The error term holds uniformly for all β in a compact set.

This is the same result as Skovgaard (1981). It also gives the same result as the second method in Daniels (1983), but that was only derived for $p = 1$. The expansion is transformation invariant, or more correctly equivariant, that is that if the parameter is transformed the approximation is transformed in the same way as the density, a property the Edgeworth expansion does not share. The value of α will be transformed, but the value of θ^* is invariant under transformations of β. The determination of α has a clear geometrical interpretation, because α is the maximum likelihood estimator for γ if the value $\tau(\theta(\beta))$ is observed in the exponential subfamily given by $\theta_0 + d\theta/d\beta\ \gamma$. It follows that if $\theta(\beta)$ is in the interior of D there exists a solution α and it is essentially unique, that is if there are two solutions they represent the same probability measure.

For nonlinear regression the results simplify even more. Let the k-dimen-

sional Y be normally distributed with mean vector $\eta(\beta)$, which is a known function of β and variance matrix $\sigma^2 I$. Let $\dot{\eta}$ be the matrix of derivatives of η and $\ddot{\eta}$ the array of second derivatives. We assume that σ^2 is known.

Theorem 2 Nonlinear regression

Under regularity conditions for $\eta(\beta)$ the density of $\hat{\beta}$ at β can be expanded as

$$p_n(\beta) = \left(\frac{n}{2\pi\sigma^2}\right)^{p/2} \exp\{-\frac{n}{2}\sigma^{-2}(\eta - \eta_0)'P(\eta - \eta_0)\}$$

$$\{|\dot{\eta}'\dot{\eta}|^{-\frac{1}{2}}|\dot{\eta}'\dot{\eta} + (\eta - \eta_0)'(I - P)\ddot{\eta}| + o(1/n)\} ,$$

where P is the projection $\dot{\eta}(\dot{\eta}'\dot{\eta})^{-1}\dot{\eta}'$. The error term holds uniformly for all β in a compact set.

3. REGULARITY CONDITIONS

In order to prove his theorem, Field (1982) lists a number of regularity conditions, which we now will comment upon. His Assumption 1 is that the estimating equation has a unique solution. Skovgaard (1981) has an interesting way of overcoming this problem. Instead of the density of the maximum likelihood estimator his expansion approximates the intensity of the point process of local maxima of the likelihood function. Assumption 2 requires that the joint density of $(\Sigma_i \psi_1(x_i,\beta),\ldots,\Sigma_i \psi_p(X_i,\beta),\hat{\beta})$ exists and has Fourier transforms which are absolutely integrable under as well the true density as the conjugate density. For maximum likelihood estimation in exponential families these densities do not exist, because there is a one-to-one correspondence between the components, for the one-dimensional case the components in the mean value parametrization are $(\bar{t}-\tau,\bar{t})$. However, as long as the relation between the variables is linear the argument of Lemma 1 in Field (1982) can still be applied.

His Assumption 3 has 6 parts, which are somewhat technical. Except for parts (ii) and (iii) they are satisfied if $\theta(\beta)$ is 5 times differentiable, because we are considering maximum likelihood estimation and because the distribution of X is an exponential family. Part (ii) is satisfied if $\theta(\beta)$ is an interior point in D, which we have to assume. This is, however, also necessary for the differentiability of $\chi(\theta)$. Part (iii) requires that the two matrices, whose determinants appear in the approximation, are non-singular. $D(\theta^*)$, which in this case is

$$\left(\frac{d\theta}{d\beta}\right)' \frac{d^2\chi}{d\theta^{*2}} \frac{d\theta}{d\beta},$$

is the information about α under the hypothesis that θ is of the form $\theta_0 + d\theta/d\beta\ \alpha$, for some α, which is an exponential subfamily. Thus it is positive definite if different values of θ give different probability measures

and $d\theta/d\beta$ has full rank. The other matrix has to be assumed non-singular.

4. PROOFS

Theorem 1 First we need to determine α defined by the equations for $r = 1,\ldots,p$

$$\int \psi_r(x,\beta) \exp\{\Sigma_j \alpha_j \psi_j(x,\beta)\} f(x)\,dx = 0 ,$$

or

$$\int \frac{d}{d\alpha_r} \exp\{\Sigma_j \alpha_j \psi_j(x,\beta)\} f(x)\,dx = 0 .$$

The left hand side is, after insertion of the expressions for ψ_j and $f(x)$

$$\int \frac{d}{d\alpha_r} \exp[\{t(x) - \tau(\theta)\}' \frac{d\theta}{d\beta}\alpha + t(x)'\theta_0]\mu(dx)/\phi(\theta_0)$$

Because it is an exponential family we can interchange integration and differentiation and calculate the integral. We find

$$\frac{d}{d\alpha_r}[\exp\{-\tau(\theta)' \frac{d\theta}{d\beta}\alpha\}\, \phi(\theta_0 + \frac{d\theta}{d\beta}\alpha)/\phi(\theta_0)] .$$

This derivative is zero iff the derivative of the logarithm is zero and that is

$$\{\tau(\theta_0 + \frac{d\theta}{d\beta}\alpha) - \tau(\theta(\beta))\}' \frac{d\theta}{d\beta_r} ,$$

thus giving the equation for α.

The rest of the proof is simple, because the integrals can be calculated using the conjugate distribution, which is the distribution with parameter $\theta^* = \theta_0 + d\theta/d\beta\,\alpha$. We list the values

$$C^{-1} = \int \exp\{\Sigma_j \alpha_j \psi_j(x,\beta)\} f(x)\,dx$$

$$= \exp\{-\tau(\theta)' \frac{d\theta}{d\beta}\alpha + \chi(\theta^*) - \chi(\theta_0)\}$$

$$A = \{E_{\theta^*} \partial\psi_i(x,\beta)/\partial\beta_r\}_{1\leq i,r\leq p}$$

$$= \{\tau(\theta^*)-\tau(\theta)\}' \frac{d^2\theta}{d\beta^2} - \frac{d\theta'}{d\beta} \frac{d^2\chi}{d\theta^2} \frac{d\theta}{d\beta}$$

$$\Sigma = \{E_{\theta^*} \psi_i(x,\beta)\psi_r(x,\beta)\}_{1\leq i,r\leq p}$$

$$= \frac{d\theta'}{d\beta} \frac{d^2\chi}{d\theta^{*2}} \frac{d\theta}{d\beta} .$$

In the latter expression a term vanishes because of the definition of θ^*. Insertion in Fields formula then yields the result.

Theorem 2 For the nonlinear regression $\theta(\beta) = \eta(\beta)$ and the sufficient statistic is $t = Y/\sigma^2$. In this case $\chi(\theta) = \frac{1}{2}\sigma^{-2}\Sigma\theta_i^2 = \frac{1}{2}\sigma^{-2}\eta'\eta$ and $\tau(\theta) = \sigma^{-2}\eta$, such that the equation for α is

$$\{\eta(\beta_0) + \dot{\eta}\alpha - \eta(\beta)\}'\dot{\eta} = 0$$

giving

$$\alpha = (\dot{\eta}'\dot{\eta})^{-1}\dot{\eta}'(\eta - \eta_0)$$

such that $d\theta/d\beta\alpha = \dot{\eta}\alpha$ is the projection of the discrepancy between the mean vector for the estimated value and for the true value on the tangent space in the estimated value. Using P for the projection $\dot{\eta}(\dot{\eta}'\dot{\eta})^{-1}\dot{\eta}'$ we find

$$\theta^* = \eta_0 + P(\eta - \eta_0)$$

$$C^{-1} = \exp\{-\tfrac{1}{2}\sigma^{-2}(\eta - \eta_0)'P(\eta - \eta_0)\}$$

$$A = -\sigma^{-2}\{\dot{\eta}'\dot{\eta} + (\eta - \eta_0)'(I - P)\ddot{\eta}\}$$

$$\Sigma = \sigma^{-2}\dot{\eta}'\dot{\eta} .$$

The result now follows from Theorem 1.

5. AN EXAMPLE

The Michaelis-Menten reaction describes the velocity of an enzymatic process as

$$v = V_{max}\, c/(K_m + c),$$

where V_{max} is the maximal velocity, c the concentration at which the velocity is half of maximal. From measurements at different known concentrations $c_1,\ldots,c_k$ we want to estimate and infer about the parameters K_m and V_{max}. Suppose that the logarithm of v_i is normally distributed with mean $\ell n\, V_{max} + \ell n\, c_i - \ell n(c_i + K_m)$ and variance σ^2. Because $\ell n\, V_{max}$ enters in such a simple way the marginal distribution of $\hat{K}_m$ as well as its approximations are independent of V_{max}. Figure 1 shows different approximations to the marginal density of $\hat{K}_m$ assuming a true value of 2 and with measurements at concentrations $c_i = 1,2,\ldots,6$. The variance σ^2 is chosen to 0.2^2. From a mathematical point of view it is natural to extend the parameter space, letting $K_m > -1$. Some of the approximations below have other lower limits.

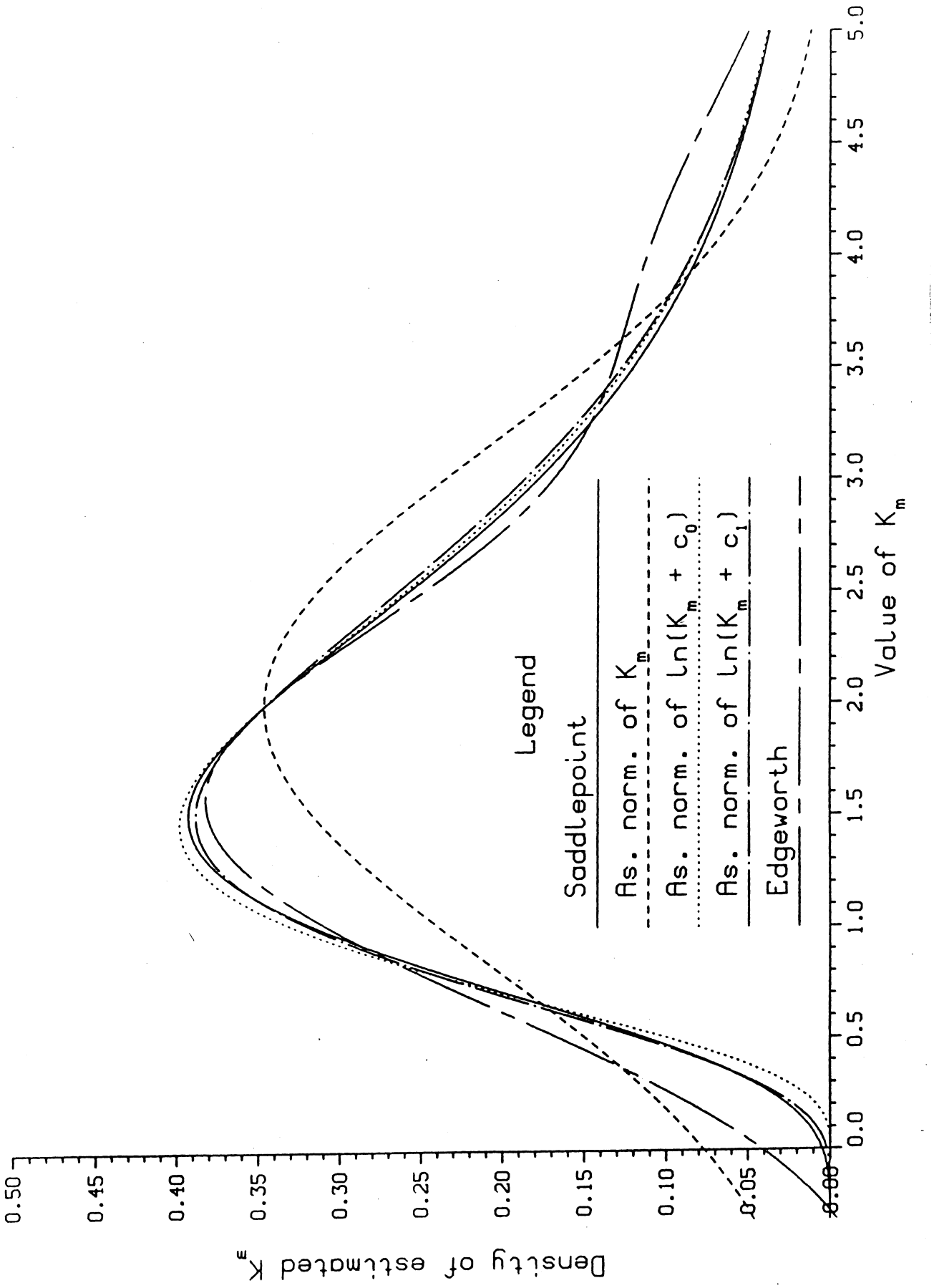
The figure shows the saddlepoint approximation described in Section 2. It is compared to the asymptotic normal distribution of $\hat{K}_m$. This approximation is poor, because of large nonlinearity. The true distribution of $\hat{K}_m$ is rather skew; the first approximation to the skewness is 1.58, cf. Hougaard (1981). The approximate normal distribution is not equivariant under nonlinear parameter transformations. Using the asymptotic normality of some function $g(\hat{K}_m)$ we can find other approximations to the density of $\hat{K}_m$. The problem is of course to choose some appropriate function g. Hougaard (1984) found a differential equation for g, such that the skewness and bias of $g(\hat{K}_m)$ asymptotically are of lower order than generally and such that the asymptotic variance of $g(\hat{K}_m)$ is constant. As this equation is fairly complicated in the present case, we must be satisfied with less, for example finding a function

g, with zero asymptotic skewness and bias in a single specified point. Following Bates and Watts (1981) and Ross (1970) we suggest to choose a concentration c_0, such that the skewness and bias of $g(\hat{K}_m) = \ell n(\hat{K}_m + c_0)$ asymptotically is zero for $K_m = 2$. With the present design this requires $c_0 = 0.18$. This approximation to the density of $\hat{K}_m$ follows the saddlepoint approximation fairly well. In practice we cannot find the optimal value of c_0, because it depends on the true unknown value of K_m. Instead we choose a value based on prior knowledge of K_m or based on the estimate. For comparison we have also examined the value $c_1 = 0.41$, which is the best for $K_m = 1.4$. For $K_m = 2$ the skewness of $\ell n(\hat{K}_m + c_1)$ is 0.15.

Also the Edgeworth expansion is shown for the distribution of $\hat{K}_m$ including the first correction term. This approximation suggests a partially negative density. In this case the approximate density is negative until $K_m = -0.30$, but the cumulative distribution function is negative until $K_m = 0.35$. Where the asymptotic normal distribution is based on simple approximations to the first and second moments, assuming that the third central moment is 0, the first correction term in the Edgeworth expansion utilizes better approximations only to the first and third moments, because the correction to the variance is of lower order. The Edgeworth expansion is not equivariant under nonlinear parameter transformations, for $\ell n(K_m + c_0)$ it coincides with the asymptotic normal distribution. The first correction term in the Edgeworth expansion depends only on the parameter effects nonlinearity. The saddlepoint approximation, however, also includes the intrinsic nonlinearity.

ACKNOWLEDGEMENTS

The comments from Søren Johansen and Ib Skovgaard are greatly appreciated.


Density of estimated K_m
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Value of K_m
Legend
SaddLepoint
As. norm. of K_m
As. norm. of $Ln(K_m + c_0)$
As. norm. of $Ln(K_m + c_1)$
Edgeworth

REFERENCES

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

Bates, D.M. & Watts, D.G. (1981). Parameter transformations for improved approximate confidence regions in nonlinear least squares. Ann. Statist. 9, 1152-1167.

Daniels, H.E. (1983). Saddlepoint approximations for estimating equations Biometrika 70, 89-96.

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

Hougaard, P. (1981). The appropriateness of the asymptotic distribution in a nonlinear regression model in relation to curvature. Research Report 81/9. Statistical Research Unit, Danish Medical and Social Science Research Councils, Copenhagen, Denmark. To appear in J.R.Statist.Soc. B 47.

Hougaard, P. (1982). Parametrizations of non-linear models. J.R.Statist.Soc. B 44, 244-252.

Hougaard, P. (1984). Parameter transformations in multiparameter nonlinear regression models. Preprint 1984/2. Institute of Mathematical Statistics, University of Copenhagen, Denmark.

Ross, G.J.S. (1970). The efficient use of function minimization in non-linear maximum-likelihood estimation. Appl. Statist. 19, 205-221.

Skovgaard, I.M. (1981). Large deviation approximations for maximum likelihood estimators. Preprint 1981/9. Institute of Mathematical Statistics, University of Copenhagen, Denmark.

PREPRINTS 1983

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

No. 1 Jacobsen, Martin: Two Operational Characterizations of Cooptional Times.

No. 2 Hald, Anders: Nicholas Bernoulli's Theorem.

No. 3 Jensen, Ernst Lykke and Rootzén, Holger: A Note on De Moivre's Limit Theorems: Easy Proofs.

No. 4 Asmussen, Søren: Conjugate Distributions and Variance Reduction in Ruin Probability Simulation.

No. 5 Rootzén, Holger: Central Limit Theory for Martingales via Random Change of Time.

No. 6 Rootzén, Holger: Extreme Value Theory for Moving Average Processes.

No. 7 Jacobsen, Martin: Birth Times, Death Times and Time Substitutions in Markov Chains.

No. 8 Hougaard, Philip: Convex Functions in Exponential Families.

PREPRINTS 1984

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

No. 1 Rootzén, Holger and Sternby, Jan: Consistency in Least Squares Estimation: A Bayesian Approach.

No. 2 Hougaard, Philip: Parameter Transformations in Multiparameter Nonlinear Regression Models.

No. 3 Jacobsen, Martin: Cooptional Times and Invariant Measures for Transient Markov Chains.

No. 4 Rootzén, Holger: Attainable Rates of Convergence of Maxima.

No. 5 Asmussen, Søren and Thorisson, Hermann: Boundary Problems and Large Deviation Results for Queue Length Processes.

No. 6 Björnsson, Ottő J.: Notes on Right-(Left-)Continuous Functions.

No. 7 Hougaard, Philip: Frailty Models Derived from the Stable Distributions.

No. 8 Hougaard, Philip: Saddlepoint Approximations for Curved Exponential Families.