Ann. Inst. Statist. Math.
Vol. 51, No. 3, 463–478 (1999)

SOME PROPERTIES AND IMPROVEMENTS OF THE SADDLEPOINT APPROXIMATION IN NONLINEAR REGRESSION

ANDREJ PÁZMAN*

Department of Probability and Statistics, Faculty of Mathematics and Physics, Comenius University, 842 15 Bratislava, Slovak Republic

(Received June 9, 1997; revised December 9, 1997)

Abstract. We summarize properties of the saddlepoint approximation of the density of the maximum likelihood estimator in nonlinear regression with normal errors: accuracy, range of validity, equivariance. We give a geometric insight into the accuracy of the saddlepoint density for finite samples. The role of the Riemannian curvature tensor in the whole investigation of the properties is demonstrated. By adding terms containing this tensor we improve the saddlepoint approximation. When this tensor is zero, or when the number of observations is large, we have pivotal, independent, and χ^2 distributed variables, like in a linear model. Consequences for experimental design or for constructions of confidence regions are discussed.

Key words and phrases: Least squares, curvatures, differential geometry, distribution of estimators, confidence regions.

1. Introduction

We consider a nonlinear regression model with normal errors

$$y = \eta(\theta) + \epsilon; \qquad \theta \in \Theta \tag{1.1}$$
$$\epsilon \sim N(0, \sigma^2 I)$$

with $y \in R^N$, and with a p-dimensional parametric space $\Theta \subset R^p$. The usual regularity assumptions are made: the mapping $\eta(\cdot)$ is one-to-one, and the matrix $J(\theta)$ with entries

$$J_{ij}(\theta) = \frac{\partial \eta_i(\theta)}{\partial \theta_j}$$

is of full rank on $\mathrm{Int}(\Theta)$.

* Supported by the project no. 1/4196/97 of the Slovak Grant Agency VEGA.

In model (1.1) neither the moments nor the probability density of the maximum likelihood estimator (the least squares estimator)

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \|y - \eta(\theta)\|^2$$

can be expressed explicitly. The familiar normal approximation gives both of them, but the level of approximation is often not sufficient. Higher-order approximations can be obtained for the moments, but they are cumbersome yet for moments of order two, and they are only "local approximations". Curiously, a much better approximation has been obtained for the probability density of the estimator, namely the so-called saddle-point (or "flat") density (abbreviated SPD in this paper). The SPD is equal to:

$$q(\hat{\theta} \mid \bar{\theta}) = \frac{\det[Q(\hat{\theta}, \bar{\theta})]}{(2\pi)^{p/2}\sigma^p \det^{1/2}[M(\hat{\theta})]} \exp\left\{-\frac{1}{2\sigma^2}\|P(\hat{\theta})[\eta(\hat{\theta}) - \eta(\bar{\theta})]\|^2\right\} \tag{1.2}$$

where $\bar{\theta}$ is the true value of θ,

$$M(\theta) = J^T(\theta)J(\theta)$$

is the Fisher information matrix for $\sigma = 1$,

$$P(\theta) = J(\theta)M^{-1}(\theta)J^T(\theta)$$

is a projector, and

$$Q_{ij}(\hat{\theta}, \theta) = M_{ij}(\hat{\theta}) + [\eta(\hat{\theta}) - \eta(\theta)]^T[I - P(\hat{\theta})]H_{ij}(\hat{\theta})$$

is a modified information matrix having a meaning by its own. We denoted

$$H_{ij}^k(\theta) = \frac{\partial^2 \eta_k(\theta)}{\partial\theta_i \partial\theta_j}.$$

The name "saddlepoint density" is justified by the fact that (1.2) can be derived by the saddlepoint technique (cf. Hougaard (1985) for this derivation and for an asymptotic justification). The saddlepoint approximation technique in general gives powerful tools for "non-local" approximations (cf. Jensen (1995)), which is for example important when the stress is on the tails of the density. However, another non-local method to obtain the SPD is the geometric method presented in Pázman (1984), where the density (1.2) has been published for the first time. An independent derivation, which in a preprint form probably influenced Hougaard (1985) has been presented in Skovgaard (1985). This approach has been extended in Skovgaard (1990) to more general models. Another geometric proof has been presented and explained in detail in Pázman (1993*b*). Later, it has been also found (cf. Pázman (1990)) that the properties of (1.2) are especially good if model (1.1) has a "zero Riemannian curvature tensor" (see below). Since such models are called "flat" in differential geometry, the approximation (1.2) may be called "flat"

as well. In this paper we prefer the geometric approach, since it gives a better insight into the approximation for a small number of observations.

The aim of this paper is to summarize some properties of the SPD. In Section 2 we present some elementary properties. In Section 3 we derive the properties of the gradient of $(1/2)\|\eta(\hat{\theta}) - \eta(\theta)\|^2$, which, under the assumption that the model is flat (or under the assumption that N is large), and under the validity of the SPD, allow to prove that the random variables $\sigma^{-2}\|P(\hat{\theta})[\eta(\hat{\theta}) - \eta(\theta)]\|^2$ and $\sigma^{-2}\|[I - P(\hat{\theta})][y - \eta(\bar{\theta})]\|^2$ are χ^2 distributed and independent, like in linear models. Interesting is also the role of the covariant derivative of the mentioned gradient as it appears in Proposition 1. The geometric insight into the SPD given in Section 4 shows how important it is for the accuracy of the SPD, that the product of the intrinsic curvature and of σ is small, i.e. that the probability given in the right-hand side of (4.2) is much smaller than 1. Under this assumption one can even improve the SPD, as shown in Section 4. This improvement consists of adding to the SPD further terms depending on the Riemannian curvature tensor. These terms disappear quickly with σ tending to zero, which explains why the SPD is asymptotically so good.

The presented investigation is essentially theoretical, in that it explains some properties of the SPD. A part of the results have been presented by the author elsewhere, spread in different papers. Here we summarize them, and we present proofs and argumentations, which are more straightforward, and we also extend some statements. In Section 6 we discuss briefly the possibilities of practical implementations. The SPD allows to "see" the properties of the estimator $\hat{\theta}$ before performing the experiment, which can be important in experimental design or in the diagnostics of the model. The derived pivotal variables allow to check the validity of the standard likelihood confidence regions, or to construct new regions. An example is given in Section 6.

2. Elementary properties of the SPD

(i) The approximate density (1.2) can be easily computed "point by point" because (1.2) contains just first and second order derivatives of $\eta(\theta)$ and requires some matrix manipulations.

(ii) The expression (1.2) is "equivariant" (in contrast to the normal approximation, which is not), i.e. to obtain (1.2) for a reparametrized model we have just to multiply the expression (1.2) by the Jacobian of the reparametrization.

(iii) The density (1.2) is exact and normal in linear models. It is also exact in intrinsically linear models, since the last can be obtained from linear models by suitable reparametrizations.

Accuracy. For large N the error term in the approximation is of order $O(N^{-1})$, while for the normal or the Edgeworth approximations it is of order $O(N^{-1/2})$ (cf. Hougaard (1985), Barndorff-Nielsen and Cox (1979)). In general, simulations demonstrated very good coincidence with simulated densities for small samples. As mentioned, for the particular case of "flat" models considered below, the SPD has especially good properties.

Range of validity. The expression (1.2) is non-negative as long as the determinant of $Q(\hat{\theta}, \bar{\theta})$ is non-negative. A sufficient condition to ensure that the matrix $Q(\hat{\theta}, \bar{\theta})$ is positive semidefinite, is the inequality

$$\|[I - P(\hat{\theta})][\eta(\hat{\theta}) - \eta(\bar{\theta})]\| \leq 1/K_{\text{int}}(\hat{\theta}). \tag{2.1}$$

Here $K_{\text{int}}(\hat{\theta})$ is the measure of intrinsic nonlinearity (= curvature) of Bates and Watts (1980) for $\sigma = 1$. (See Pázman (1993*b*) Proposition 4.2.1 for an explicit expression for $K_{\text{int}}(\theta)$. The proof of (2.1) follows from the first lines of the proof of lemma 7.1.1.b in the same reference.) In those points where (2.1) does not hold, the expression

$$\exp\left\{-\frac{1}{2\sigma^2}\|P(\hat{\theta})[\eta(\hat{\theta}) - \eta(\bar{\theta})]\|^2\right\}$$

should be negligible when compared to 1, to ensure that $q(\hat{\theta} \mid \theta)$ is close to zero. Another limitation of validity of (1.2) is due to the possible overlapping of the expectation surface of model (1.1),

$$\{\eta(\theta) : \theta \in \Theta\}$$

over itself. In terms of the expression (1.2) this means that the expression $\|P(\hat{\theta})[\eta(\hat{\theta}) - \eta(\bar{\theta})]\|$ may decrease although the distance $\|\hat{\theta} - \bar{\theta}\|$ is increasing. As a consequence we can have false peaks of the function $\hat{\theta} \to g(\hat{\theta} \mid \bar{\theta})$ given by (1.2). This is related to the fact mentioned in Skovgaard (1985), that in general the expression (1.2) is not the density but the intensity of $\hat{\theta}$. This however has practical consequences only if there is overlapping.

Flexibility. By geometrical methods the SPD can be modified, to adapt to the case of weighted least squares, or to the case of the maximum posterior estimator. In particular, this allows to take into account the probability distribution on the boundary of the parameter space Θ, etc. (cf. Pázman (1993*b*) for references).

3. Some small sample properties of the SPD

Let us consider the random vector

$$v(\hat{\theta}) = \frac{\partial}{\partial \hat{\theta}} \frac{1}{2}\|\eta(\hat{\theta}) - \eta(\bar{\theta})\|^2 = J^T(\hat{\theta})[\eta(\hat{\theta}) - \eta(\bar{\theta})].$$

As shown in Proposition 1 given below, if $\hat{\theta}$ is distributed according to the SPD (1.2), then the random vector $v(\hat{\theta})$ is "locally normal" in a certain sense. This has two statistical consequences presented below: Firstly, $v(\hat{\theta})$ is normally distributed when the information matrix $M(\theta)$ does not depend on θ. Secondly, in any model with a zero Riemannian curvature tensor (see (3.3)) the random variable

$$\xi(\hat{\theta}) = \sigma^{-2}\|P(\hat{\theta})[\eta(\hat{\theta}) - \eta(\bar{\theta})]\|^2$$

is distributed as a truncated χ^2 with p degrees of freedom (like in a linear model with a bounded parameter space). In particular, any model with $\dim(\theta) = 1$ has this property (cf. Pázman (1990)).

Geometry. As known (cf. e.g. Amari (1985)) the parameters θ correspond to the coordinates of a manifold (the coordinates are denoted by x in the classical book on differential geometry by Eisenhart (1960)). The matrix $M(\theta)$ corresponds to the Riemannian metric tensor (denoted by $g(x)$ in Eisenhart (1960)).

According to Eq. (11.3) in Eisenhart (1960), the covariant derivative of any vector function $v(\hat{\theta})$ is defined as a matrix with components

$$\left\{ \frac{Dv(\hat{\theta})}{D\hat{\theta}^T} \right\}_{ij} = \frac{\partial v_i(\hat{\theta})}{\partial \hat{\theta}_j} - v^T(\hat{\theta})\Gamma_{ij}^{\cdot}(\hat{\theta}). \tag{3.1}$$

Here $\Gamma_{ij}^{k}(\hat{\theta})$ are the Christofell symbols of the second kind as presented in Eq. (7.2) in Eisenhart (1960) (or the components of the affine connection in Amari (1985)). In case of the geometry of a nonlinear regression model with normal errors we have

$$\Gamma_{ij}^{k}(\hat{\theta}) = \sum_s \{M^{-1}(\hat{\theta})\}_{ks} J_{s\cdot}^T(\hat{\theta}) H_{ij}^{\cdot}(\hat{\theta}). \tag{3.2}$$

PROPOSITION 1. *The SPD* (1.2) *can be expressed in the form*

$$q(\hat{\theta} \mid \bar{\theta}) = \frac{\det[Dv(\hat{\theta})/D\hat{\theta}^T]}{(2\pi)^{p/2}\sigma^p \det^{1/2}[M(\hat{\theta})]} \exp\left\{ -\frac{1}{2\sigma^2} v^T(\hat{\theta}) M(\hat{\theta}) v(\hat{\theta}) \right\}.$$

PROOF. From (3.1), (3.2) we obtain after some rearrangements that $Dv(\hat{\theta})/D\hat{\theta}^T = Q(\hat{\theta}, \bar{\theta})$. To obtain the new expression in the exponent, we have just to rearrange the terms in the expression $\|P(\hat{\theta})[\eta(\hat{\theta}) - \eta(\theta)]\|^2$ using the definition of $P(\hat{\theta})$. □

COROLLARY 1. *If the information matrix $M(\theta)$ does not depend on θ (i.e. $M(\theta) = M$), then the covariant derivative is an ordinary derivative (cf. Eisenhart* (1960), *Section* I.11). *Hence from Proposition* 1 *we obtain directly that in this case $v(\hat{\theta})$ is distributed normally $N(0, \sigma^2 M)$ on $\mathrm{int}(\Theta)$. Since the range of $v(\hat{\theta})$ is bounded, one can say that $v(\hat{\theta})$ has a "truncated" normal distribution.*

The Riemannian curvature tensor. In the Euclidean geometry, which corresponds to model (1.1), the Riemannian curvature tensor (cf. Eisenhart (1960), Section 1.8) is a 4-dimensional tensor $R(\theta)$ with components:

$$R_{hijk}(\theta) = H_{hj}^{\cdot}(\theta)^T [I - P(\theta)] H_{ik}^{\cdot}(\theta) - H_{hk}^{\cdot}(\theta)^T [I - P(\theta)] H_{ij}^{\cdot}(\theta). \tag{3.3}$$

(See Section 1 for the definitions of $H(\theta)$ and $P(\theta)$.) There is no direct statistical interpretation of $R(\theta)$, and also its geometric interpretation is not quite direct.

However, it is fundamental to the Riemannian geometry, and we shall show below that it is closely related to the properties of the SPD. We note that $R(\theta)$ is of some interest also for the normal approximation of the density of $\hat{\theta}$, since $R(\theta) \equiv 0$ (identically) implies that there is a reparametrization of model (1.1) making the asymptotic variance matrix of the new parameters constant (cf. Hougaard (1986)).

COROLLARY 2. *If $\hat{\theta}$ is distributed according to* (1.2) *and $R(\theta) \equiv 0$, then $\xi(\hat{\theta})$ is distributed as a "truncated" χ^2 variable with p degrees of freedom.*

PROOF. According to a result of Riemann (cf. Eisenhart (1960), Section I.9.), the condition $R(\theta) \equiv 0$ is necessary and sufficient for the existence of a reparametrization $\beta = \beta(\theta)$ such that $M(\beta) = M$ does not depend on β. Here by $M(\beta)$ we denoted the information matrix in the reparametrized model $y = \nu(\beta) + \epsilon$, with $\nu(\beta) = \eta[\theta(\beta)]$, and by $v(\hat{\beta})$, and $P(\beta)$, etc. we denote other expressions corresponding to this new model. We can apply Corollary 1, which implies that $v^T(\hat{\beta})M^{-1}v(\hat{\beta})$ is distributed χ^2 with p degrees of freedom. Further we have evidently

$$\zeta(\hat{\theta}) = \sigma^{-2}\|P(\hat{\theta})[\eta(\hat{\theta}) - \eta(\bar{\theta})]\|^2 = \sigma^{-2}\|P(\hat{\beta})[\nu(\hat{\beta}) - \nu(\bar{\beta})]\|^2 = v^T(\hat{\beta})M^{-1}v(\hat{\beta})$$

and to finish the proof we use that $v(\hat{\beta})$ is distributed $N(0, M)$. However $\zeta(\hat{\theta})$ may have a truncated χ^2 distribution, since in the general case, the set Θ may not be equal to R^p. □

4. The geometric insight into the small-sample accuracy of the SPD and the improvements of the SPD

Let us denote by $\omega^{(1)}, \ldots, \omega^{(N-p)}$ an orthonormal set of vectors, which are also orthogonal to the expectation surface at the point $\hat{\theta}$, i.e.

$$P(\hat{\theta})\omega^{(i)}(\hat{\theta}) = 0; \qquad i = 1, \ldots, N - p.$$

Since $\hat{\theta}(y)$ is defined uniquely with probability one, one can introduce new coordinates $(\hat{\theta}, b)$ of y where

$$b_i = b_i(y) = [y - \eta(\bar{\theta})]^T\omega^{(i)}(\hat{\theta}(y)); \qquad i = 1, \ldots, N - p.$$

We have

$$y - \eta(\bar{\theta}) = \sum_i b_i\omega^{(i)}(\hat{\theta}) + P(\hat{\theta})[\eta(\hat{\theta}) - \eta(\bar{\theta})]$$

since $P(\hat{\theta})[y - \eta(\bar{\theta})] = P(\hat{\theta})[\eta(\hat{\theta}) - \eta(\bar{\theta})]$. Hence, according to the Pythagorean relation, in terms of the new coordinates one can factorize the density of y as follows

$$f(y \mid \bar{\theta}) \mid_{y=y(\hat{\theta},b)} = \text{const} \exp\{-\|b\|^2/(2\sigma^2)\} \exp\{-\|P(\hat{\theta})[\eta(\hat{\theta}) - \eta(\bar{\theta})]\|^2/(2\sigma^2)\}.$$

Evidently, if b would not depend on $\bar{\theta}$, this factorization would demonstrate the sufficiency of $\hat{\theta}$. This is not the case, however, still the dependence of b on $\bar{\theta}$ is sufficiently weak to make the new coordinates useful.

For further use we need the Jacobian of the mapping $y \to (\hat{\theta}, b)$. A direct derivation is presented in Pázman (1993*b*), Proposition 7.1.1b, but essentially the result must be similar as in Barndorff-Nielsen (1980), who, for another purpose, used the approximate ancillary statistics $a_i(y) = [y - \eta(\hat{\theta})]^T \omega^{(i)}(\hat{\theta})$ instead of $b_i(y)$. The Jacobian is equal to

$$\left|\det\left(\frac{\partial y}{\partial \hat{\theta}^T}, \frac{\partial y}{\partial b^T}\right)\right| = \text{const}\, \frac{|\det(\text{observed inform. matrix})_{\hat{\theta}=\theta}|}{\det^{1/2}(\text{expected inform. matrix})_{\hat{\theta}=\theta}}.$$

Since $[y - \eta(\hat{\theta})] = [I - P(\hat{\theta})][y - \eta(\hat{\theta})]$, the observed information matrix can be written in the form

$$\begin{aligned} -\frac{\partial^2 \ln f(y \mid \theta)}{\partial \theta_i \partial \theta_j}\bigg|_{\theta=\hat{\theta}} &= \sigma^2\{M(\hat{\theta}) + [\eta(\hat{\theta}) - y]^T H_{ij}(\hat{\theta})\} \\ &= \sigma^{-2}[Q(\hat{\theta}, \bar{\theta}) + D(b, \hat{\theta})]_{ij} \end{aligned}$$

where

$$D_{ij}(b, \hat{\theta}) = -\sum_{k=1}^{N-p} b_k [\omega^{(k)}(\hat{\theta})]^T H_{ij}(\hat{\theta}).$$

The joint density of $(\hat{\theta}, b)$ is then

$$p_{\bar{\theta}}(\hat{\theta}, b) = f(y \mid \bar{\theta})\,|_{y=y(\hat{\theta},b)} \times \text{Jacobian}$$

and the exact density of $\hat{\theta}$ (or more precisely the exact "intensity" if there is an overlapping) is obtained after some rearrangements

$$q_{exact}(\hat{\theta} \mid \bar{\theta}) = \int_{\mathcal{D}(\hat{\theta})} p_{\bar{\theta}}(\hat{\theta}, b) db = q(\hat{\theta} \mid \bar{\theta}) I(\hat{\theta}, \bar{\theta})$$

where $q(\hat{\theta} \mid \bar{\theta})$ is the SPD, where

$$I(\hat{\theta}, \bar{\theta}) = \det{}^{-1}[Q(\hat{\theta}, \bar{\theta})] \int_{\mathcal{D}(\hat{\theta})} \det[Q(\hat{\theta}, \bar{\theta}) + D(b, \hat{\theta})] \phi(b) db$$

and where $\phi(b)$ denotes the $N - p$ dimensional normal density $N(0, \sigma^2 I)$. The range of integration, $\mathcal{D}(\hat{\theta})$, is given by

$$\mathcal{D}(\hat{\theta}) = \{b : Q(\hat{\theta}, \bar{\theta}) + D(b, \hat{\theta}) \text{ is positive definite}\}$$

since a solution of $\partial \|y - \eta(\theta)\|^2 / \partial \theta = 0$ is a minimum iff $\partial^2 \|y - \eta(\theta)\|^2 / \partial \theta \partial \theta^T$ ($\approx$ the observed information matrix) is positive definite. The shape of the set $\mathcal{D}(\hat{\theta})$

may be very complicated. What is proposed here, is to approximate $\mathcal{D}(\hat{\theta})$ simply by R^{N-p}, i.e. to take

$$I(\hat{\theta},\bar{\theta}) \doteq \det{}^{-1}[Q(\hat{\theta},\bar{\theta})]E\{\det[Q(\hat{\theta},\bar{\theta}) + D(b,\hat{\theta})]\} \tag{4.1}$$

where E denotes the mean with respect to $\phi(b)$, and $\hat{\theta}$ is fixed.

To justify this approximation we start from the fact that for any solution θ^* of the equation $\partial\|y-\eta(\theta)\|^2/\partial\theta = 0$, such that $\|y-\eta(\theta^*)\|^2 < 1/K_{\text{int}}(\theta^*)$ (= the radius of curvature), the observed information matrix at $\theta = \theta^*$ is positive definite. On the other hand we have evidently

$$\begin{aligned} P_{\bar{\theta}}\{y : \|y-\eta(\hat{\theta})\| \geq 1/K_{\text{int}}(\hat{\theta})\} &\leq P_{\bar{\theta}}\{y : \|y-\eta(\bar{\theta})\| \geq 1/K_{\text{int}}(\hat{\theta})\} \\ &= \Pr\{\chi^2_{N-p} \geq 1/[\sigma^2 K^2_{\text{int}}(\hat{\theta})]\} \end{aligned} \tag{4.2}$$

and this probability can be neglected if the product $\sigma \cdot K_{\text{int}}(\hat{\theta})$ is "sufficiently small". Then the approximation (4.1) can be applied.

To compute the mean in (4.1) we use the decomposition of the determinant (we omit to write the symbols $\hat{\theta}$ and $\bar{\theta}$):

$$\det[Q + D(b)] = \det[Q] + \sum_{s=1}^{p} \sum_{U \in J_s} \det[K(U)].$$

Here J_s is the set of all s-point subsets of $\{1,\dots,p\}$, and $K(U)$ is the $p \times p$ matrix with the i-th column equal either to the i-th column of $D(b)$ (if $i \in U$), or to the i-th column of Q (if $i \notin U$). Using now the Laplace decomposition of $\det[K(U)]$ over the columns corresponding to U we obtain

$$\det[K(U)] = \sum_{V \in J_s} \pm \det[D_{U,V}(b)] \det[Q_{U^*V^*}]$$

where $U^* = J_s \backslash U$, and where $D_{U,V}(b)$ is a submatrix of $D(b)$ with row and column subscripts taken from U and V.

Now if $U, V \in J_s$ and s is odd, then $E\{\det[D_{U,V}(b)]\} = 0$, since $D(b)$ is linear in b. On the other hand, if s is even, then $E\{\det[D_{U,V}(b)]\}$ is a polynomial in the components of R and Q. Indeed, $D(b)$ is linear in b, and $\det[D(b)]$ is a homogeneous polynomial in the components of $D(b)$. Hence the mean $E\{\det[D_{U,V}(b)]\}$ is a linear combination of moments of $D(b)$ of order s. But the components of $D(b)$ are normal variables, hence $E\{\det[D_{U,V}(b)]\}$ can be expressed through the second moments, more exactly through terms like

$$E\left\{\det\begin{pmatrix} D_{ik}(b), D_{il}(b) \\ D_{jk}(b), D_{jl}(b) \end{pmatrix}\right\} = \sigma^2 R_{ijkl}.$$

In the last equality we used the fact that

$$\sum_{i=1}^{N-p} \omega^{(i)}[\omega^{(j)}]^T = I - P$$

(cf. Pázman (1993*a*) for technical details). So we obtained the following theorem.

THEOREM 1. *If the probability* (4.2) *can be neglected, then the exact probability density of* $\hat{\theta}$ *is expressed in the form*

$$q_{exact}(\hat{\theta} \mid \bar{\theta}) = q(\hat{\theta} \mid \bar{\theta}) \times (\textit{polynomial in the components of } Q(\hat{\theta}, \bar{\theta}) \textit{ and of } \sigma^2 R(\hat{\theta})).$$

The absolute term of this polynomial is equal to 1.

COROLLARY 3. *If* $R(\theta) \equiv 0$ *and the probability* (4.2) *can be neglected, then the SPD* $q(\hat{\theta} \mid \bar{\theta})$ *is the exact density (called "almost exact" in Pázman* (1993*b*) *because the probability* (4.2) *is not zero). In particular* $R(\theta) = 0$ *in any model with* $\dim(\theta) = 1$, *or e.g. in the two-dimensional classical Michaelis-Menten regression model.*

COROLLARY 4. *The SPD can be improved in models with* $R(\theta) \not\equiv 0$ *by adding further terms of the polynomial. In particular, for* $\dim(\theta) = 2$ *we have*

$$q_{exact}(\hat{\theta} \mid \bar{\theta}) = \frac{\det[Q(\hat{\theta}, \bar{\theta})] + \sigma^2 R_{1212}(\hat{\theta})}{2\pi\sigma^2 \det^{1/2}[M(\hat{\theta})]} \exp\left\{-\frac{1}{2\sigma^2}\|P(\hat{\theta})[\eta(\hat{\theta}) - \eta(\bar{\theta})]\|^2\right\}.$$

If $\dim(\theta) = 3$ *we have*

$$q_{exact}(\hat{\theta} \mid \bar{\theta}) = \frac{\det[Q(\hat{\theta}, \bar{\theta})] + \sigma^2 \sum_{i,j=1}^{3} (-1)^{i+j} Q_{ij}(\hat{\theta}) R_{i+1,i+2,j+1,j+2}(\hat{\theta})}{(2\pi)^{3/2}\sigma^3 \det^{1/2}[M(\hat{\theta})]}$$
$$\times \exp\left\{-\frac{1}{2\sigma^2}\|P(\hat{\theta})[\eta(\hat{\theta}) - \eta(\theta)]\|^2\right\}$$

where the sums are modulo 3. (*Cf. Pázman* (1993*a*), *Eq.* (2.5) *for the case that* $\dim(\theta) > 3$. *Notice, that we have to put* $r = \infty$ *into the formulae given there, since we suppose here that the probability* (4.2) *can be neglected.*)

5. The asymptotic accuracy of the SPD, and pivotal variables

An analogy to the Edgeworth expansion. Since the terms of the polynomial in Theorem 1 are successively of order $O(1)$, $O(\sigma^2)$, $O(\sigma^4)$, etc., asymptotically (for replicated experiments, i.e. for $\sigma \to 0$) one obtains that in the general model (1.1)

$$q_{exact}(\hat{\theta} \mid \bar{\theta}) = q(\hat{\theta} \mid \bar{\theta})[1 + O(\sigma^2)].$$

So the relative error of approximation is $O(N^{-1})$ as obtained by the saddle-point method. We can obtain also higher order aproximations if we take into account further terms in the polynomial. Notice that such higher order improvements are related to the SPD like the Edgeworth expansion is related to the normal approximation. However, while in the Edgeworth expansion the coefficients in the correcting terms are rational functions of the moments of the exact density, here

these coefficients are polynomials of the components of the generalized information matrix $Q(\hat{\theta}, \bar{\theta})$ and of the Riemannian curvature tensor $R(\hat{\theta})$.

A pivotal variable. In nonlinear models one use for the estimation of σ^2 the maximum likelihood estimator

$$s^2 = \|y - \eta(\hat{\theta})\|^2/(N-p).$$

It is well known that it is consistent. However, in intrinsically nonlinear models it depends statistically on $\hat{\theta}$, and it is not distributed χ^2, hence it is biased for finite N.

Here we shall consider alternatively the properties of another random variable

$$\psi_{\bar{\theta}}(y) = \|[I - P(\hat{\theta})][y - \eta(\hat{\theta})]\|^2/(N-p) = \|b\|^2/(N-p)$$

which is equal to s^2 in intrinsically linear models, but which maintains the "linear" properties of s^2 also in intrinsically nonlinear models. This is formulated in the following theorem, which extends the properties of $\psi_{\bar{\theta}}(y)$ presented in Pázman (1991, 1993*a*). We present a proof which is simpler than the earlier proof.

THEOREM 2. *If $R(\theta) \equiv 0$ and the probabilities* (4.2) *can be neglected, then*

i) *$\psi_{\bar{\theta}}(y)$ and $\hat{\theta}$ are independent random variables,*

ii) *$(N-p)\psi_{\bar{\theta}}(y)/\sigma^2$ is distributed χ^2 with $N-p$ degrees of freedom,*

iii) *the mean of $\psi_{\bar{\theta}}(y)$ is equal to σ^2.*

If $R(\theta)$ is arbitrary, then the conditional cumulative distribution function of $(N-p)\psi_{\bar{\theta}}(y)/\sigma^2$ given $\hat{\theta}$ is the χ^2_{N-p} cumulative distribution function, plus a term of order $O(\sigma^2)$.

PROOF. Let us consider the conditional density $h_{\bar{\theta}}(b \mid \hat{\theta})$ of the variable b with components $b_i(y)$. We have

$$h_{\bar{\theta}}(b \mid \hat{\theta}) = p_{\bar{\theta}}(\hat{\theta}, b)/q_{exact}(\hat{\theta} \mid \bar{\theta}).$$

Using results of Section 3 we obtain

$$h_{\bar{\theta}}(b \mid \hat{\theta}) = \det[Q(\hat{\theta}, \bar{\theta}) + D(b, \hat{\theta})]\phi(b) \Big/ \int_{\mathcal{D}(\hat{\theta})} \det[Q(\hat{\theta}, \bar{\theta}) + D(b, \hat{\theta})]\phi(b)db.$$

So, supposing that the probability (4.2) is negligible, the conditional c.d.f. of $(N-p)\psi_{\bar{\theta}}(y) = \|b\|^2$, given $\hat{\theta}$, is equal to

$$\begin{aligned} F(x) &= P_{\hat{\theta}}\{y : \|b(y)\|^2 < x \mid \hat{\theta}\} \\ &= \int_{\|b\|^2 < x} \det[Q + D(b)]\phi(b)db \Big/ \int_{\|b\|^2 < \infty} \det[Q + D(b)]\phi(b)db. \end{aligned}$$

The second of these integrals has been expressed in Section 3, and the first one can be handled exactly in the same way. Just realize that it can be written as

the mean $E^*\{\det[Q + D(b)]\}$ with respect to the density $\phi(b)$ restricted to the set $\{b : \|b\|^2 < x\}$, instead of the mean with respect to the non-restricted density $\phi(b)$. By the arguments of Section 3 applied to both integrals we obtain when $R(\theta) \equiv 0$

$$F(x) = \int_{\|b\|^2<x} \phi(b)db$$

i.e. $(N-p)\psi_{\bar{\theta}}(y)/\sigma^2$ is distributed χ^2_{N-p} independently of $\hat{\theta}$ (cf. Pázman (1993*a*) for technical details). When $R(\theta)$ is not zero, one obtains

$$F(x) = \left[\det[Q] \int_{\|b\|^2<x} \phi(b)db + O(\sigma^2)\right] / [\det[Q] + O(\sigma^2)]$$

which is the χ^2_{N-p} c.d.f. up to the term $O(\sigma^2)$. □

Confidence regions. Unfortunately, despite of its good statistical properties, similar to the properties of s^2 in linear models, the variable $\psi_{\bar{\theta}}(y)$ is not an estimator of σ^2, since it depends on $\bar{\theta}$. However, as a pivotal variable it can be used, at least in principle, for the construction of confidence regions, where it is a substitute for an estimator of σ. Indeed, a corollary of Theorem 2 is that, when the probability (4.2) can be neglected, the random variable

$$\tau_{\bar{\theta}}(y) = \frac{\|P(\hat{\theta})[\eta(\hat{\theta}) - \eta(\bar{\theta})]\|^2(N-p)}{\|[I - P(\hat{\theta})][y - \eta(\bar{\theta})]\|^2 p}$$

is distributed $F_{p,N-p}$ (exactly if $R(\theta) \equiv 0$, or approximately if $R(\theta) \not\equiv 0$ and σ is small). This allows a construction of a confidence region for θ

$$\{\theta \in \Theta : \tau_\theta(y) < F_{p,N-p}(1-\alpha)\} \tag{5.1}$$

where $F_{p,N-p}(1-\alpha)$ is the $1-\alpha$ quantile of the $F_{p,N-p}$ distribution. However, this region may be incorrect when overlapping occurs. To avoid it, one can put an additional condition to obtain the region

$$\{\theta \in \Theta : \tau_{\bar{\theta}}(y) < F_{p,N-p}(1-\alpha) \text{ and } \|\eta(\hat{\theta}) - \eta(\theta)\| < K_{\text{int}}(\hat{\theta})\} \tag{5.2}$$

which has been presented in Pázman (1991). However, to use (5.2) one has to be sure that the probability (4.2) can be neglected. This may be difficult to verify, since one has no good estimator of σ in highly curved models. But in the general case, one can relatively easily compare (5.1) with the standard likelihood region, and if the two regions differ very much, one can be sure that the likelihood region is wrong, as discussed in the example in Section 6. The region (5.2) may still be correct, but at least some approximate verification of (4.2) is necessary.

6. Discussion

Although the presented paper is essentially theoretical, the obtained results have some practical consequences.

a) *The use of the SPD for the comparison (design) of experiments*

Once the experiment is performed (i.e. y is known), one has the likelihood function for inference, which is much easier to compute than the density of $\hat{\theta}$. However, in problems like the design of experiments we do not know y beforehand. Then experiments, which use least squares, are to be compared according to the distribution of $\hat{\theta}$. The covariance or the mean square error matrices of $\hat{\theta}$ would be certainly useful here, but there are no simple and efficient approximations of them available. (Cf. Clarke (1980) to see how complicated a second order approximation can be.) Therefore, instead of using moments, the solution here is the use of the approximate density of $\hat{\theta}$. The graphical presentation of the SPD or its improvement has priority here. This can be done without difficulties when $\dim(\theta) \leq 2$. When $\dim(\theta) \geq 3$, one has to plot the crossections of $q(\hat{\theta} \mid \theta)$: one fixes all components of $\hat{\theta}$ but one or two, and plots $q(\hat{\theta} \mid \theta)$ as a function of the free components. Up to a norming factor, this is the conditional density of the free components under the condition that the values of the other components are given.

Another use of $q(\hat{\theta} \mid \bar{\theta})$ in experimental design is to express explicitely the dependence of $q(\hat{\theta} \mid \bar{\theta})$ on the experimental design, and to compare experiments according to the value of the mean square error

$$\int_{\Theta} q(\hat{\theta} \mid \bar{\theta}) \|\hat{\theta} - \bar{\theta}\|^2 d\hat{\theta}.$$

One can also compute iteratively the optimum design minimizing this expression (cf. Pázman and Pronzato (1992) for details and a numerical example).

b) *Confidence regions for θ*

When the normal approximation of the density of $\hat{\theta}$ can be accepted, the (approximate) confidence ellipsoid or "linear confidence region"

$$\{\theta : (\theta - \hat{\theta})^T M(\hat{\theta})(\theta - \hat{\theta}) \leq ps^2 F_{p,N-p}(1-\alpha)\} \tag{6.1}$$

is correct. However, e.g. in Bates and Watts ((1988), p. 65) the authors "... warn the reader that linear approximate regions can be extremely misleading". In general, much better are the likelihood regions (cf. Bates and Watts (1988), Chapter 6.1.1) $\{\theta : \|y - \eta(\theta)\|^2 - \|y - \eta(\hat{\theta})\|^2 \leq ps^2 F_{p,N-p}(1-\alpha)\}$, which for intrinsically linear models can be written in the form

$$\{\theta : \|\eta(\theta) - \eta(\hat{\theta})\| \leq ps^2 F_{p,N-p}(1-\alpha)\}. \tag{6.2}$$

While both regions coincide and are exact in linear models, the likelihood region remains exact also in nonlinear models, which are still intrinsically linear. (By "exact" we mean that the confidence level is exactly $(1-\alpha)$.) In order models the likelihood region is only approximate as well. The regions (5.1) or (5.2) given in Section 5 coincide with the likelihood regions when the model is at least intrinsically linear, but they remain to be "almost exact" in a larger class of models, namely with $R(\theta) \equiv 0$ and with a negligible probability (4.2). To evaluate (4.2)

one needs an estimate of σ^2, but in the case that the model is intrinsically curved, the estimator s^2 may be wrong. What still simply can be done in particular experiments, is to compare the region (5.1) with the likelihood region (5.2).

The numerical computation of the contour of the region (5.1) is essentially no more difficult than for the likelihood region (6.2). Notice that $P(\hat{\theta})$ in (5.1) is fixed, so what is changing along the contour is just the expression $\eta(\theta)$, which appears in both regions. In the example below we present a simple technique how to obtain points of this contour numerically. When we only want to check the correctness of (6.2), it is sufficient to make the computation just for a few points.

Example. Consider the observation of the biochemical oxygen demand (BOD) discussed largely in Bates and Watts (1988), p. 41, with observed data given on p. 270, the L.S. estimator $\hat{\theta}$ and s^2 on p. 51, the "linear" confidence ellipsoids plotted on p. 55, the likelihood regions on p. 64, 202, 211. The response in the model

$$\eta(\theta, x) = \theta_1(1 - \exp\{-\theta_2 x\})$$

has been observed independently in $N = 6$ points

$$x_1 = 1, \quad x_2 = 2, \quad x_3 = 3, \quad x_4 = 4, \quad x_5 = 5, \quad x_6 = 7$$

with the results

$$y_1 = 8.3, \quad y_2 = 10.2, \quad y_3 = 19.0, \quad y_4 = 16.0, \quad y_5 = 15.6, \quad y_6 = 19,8.$$

The L.S. estimator is

$$\hat{\theta}_1 = 19.143, \qquad \hat{\theta}_2 = 0.5311$$

and

$$s^2 = 6.498.$$

We note that a similar example is considered in Seber and Wild ((1989), examples 3 and 4, p. 111), but with different data. In both cases the model gives likelihood regions which are far from the confidence ellipsoids, and which can even have infinite boundaries. This can be explained by a large parameter effect curvature of the model. Here we compare these regions with (5.2), and the difference may be explained by the intrinsic curvature.

Denote

$$\hat{z} = \exp\{-\hat{\theta}_2\} = 0.5879578,$$
$$v = (\hat{z}, \hat{z}^2, \hat{z}^3, \hat{z}^4, \hat{z}^5, \hat{z}^7)^T$$

and

$$w = (\hat{z}, 2\hat{z}^2, 3\hat{z}^3, 4\hat{z}^4, 5\hat{z}^5, 7\hat{z}^7)^T.$$

We can write

$$\eta(\hat{\theta}) = \hat{\theta}_1(\mathbf{1} - v), \qquad \frac{\partial \eta(\hat{\theta})}{\partial \theta_2} = \hat{\theta}_1 w, \qquad \frac{\partial \eta(\hat{\theta})}{\partial \theta_1} = \mathbf{1} - v.$$

This allows to obtain directly the 2×2 matrix $M(\hat{\theta})$ and the 6×6 matrix $P(\hat{\theta})$. To obtain points of the contour of (5.1) we use the polar coordinates ρ, ϕ, centered at $\hat{\theta}$ (spherical coordinates when $\dim(\theta) > 2$):

$$\theta - \hat{\theta} = \rho(\cos\phi, \sin\phi)^T.$$

We fix the angle ϕ and specify ρ which is on the contour of (5.1), i.e. which is the solution of

$$\tau(\rho) \equiv \frac{2h(\rho)}{[\|y - t(\rho)\|^2 - h(\rho)]} = F_{2,4}(1-\alpha) \tag{6.3}$$

where

$$t_i(\rho) \equiv \rho\cos\phi(1 - e^{-x_i\rho\sin\phi}); \qquad i = 1, 2, \ldots, 6$$

and

$$h(\rho) \equiv [y - t(\rho)]^T P(\hat{\theta})[y - t(\rho)].$$

The simplest way is to plot the left hand side of (6.3) as a function of ρ (see Fig. 1). This allows to omit the verification of the restriction $\|\eta(\hat{\theta}) - \eta(\theta)\| < K_{\text{int}}(\hat{\theta})$ in (5.2) since we simply take as the solution of (6.2) the first passage of the graph through the level $F_{2,4}(1-\alpha)$.

To compare, we plot also the graph of the function

$$\lambda(\rho) \equiv \|\eta(\bar{\theta}) - t(\rho)\|^2/(2s^2).$$

The first passage of this graph through the level $F_{2,4}(1-\alpha)$ gives the point of the contour of the likelihood region (6.2).

These graphs given in Fig. 1 have been obtained with the help of S-plus for 3 different values of the angle ϕ. The full lines correspond to $\lambda(\rho)$, the dotted lines to $\tau(\rho)$. The horizontal lines indicate the levels $F_{2,4}(0.9) = 4.32$ and $F_{2,4}(0.95) = 6.94$. When $\phi = 1.1$ and $(1-\alpha) = 90\%$, the point of the contour is common for (5.2) and (6.2), which does not hold when $(1-\alpha) = 95\%$. When $\phi = 2.2$ the regions have different points of contour even for $(1-\alpha) = 90\%$. For $(1-\alpha) = 95\%$ the region (5.2) is infinite, whilst the region (6.2) is finite, but for larger $(1-\alpha)$ it would be infinite as well. Further intersections of the graphs with the horizontal lines come from overlapping, and can not be taken into account. One can conclude that the likelihood region on the 95% confidence level can not be considered being reliable in the direction $\phi = 2.2$. Finally when $\phi = 6.25$ both confidence regions are infinite (for both levels of confidence).

Drawing those figures for 20 angles ϕ (which is a rather quick procedure in S-plus) we obtained that for each ϕ the confidence region (5.1), resp (5.2) is equal or larger than the likelihood region. On the 95% level there is a difference between the two regions when $0.4\pi \leq \phi \leq 0.8\pi$, and $1.98\pi \leq \phi \leq 1.998\pi$. For any other ϕ there is a coincidence of both regions. This means that the likelihood confidence region is slightly too optimistic, but up to small intervals of the angle ϕ, the

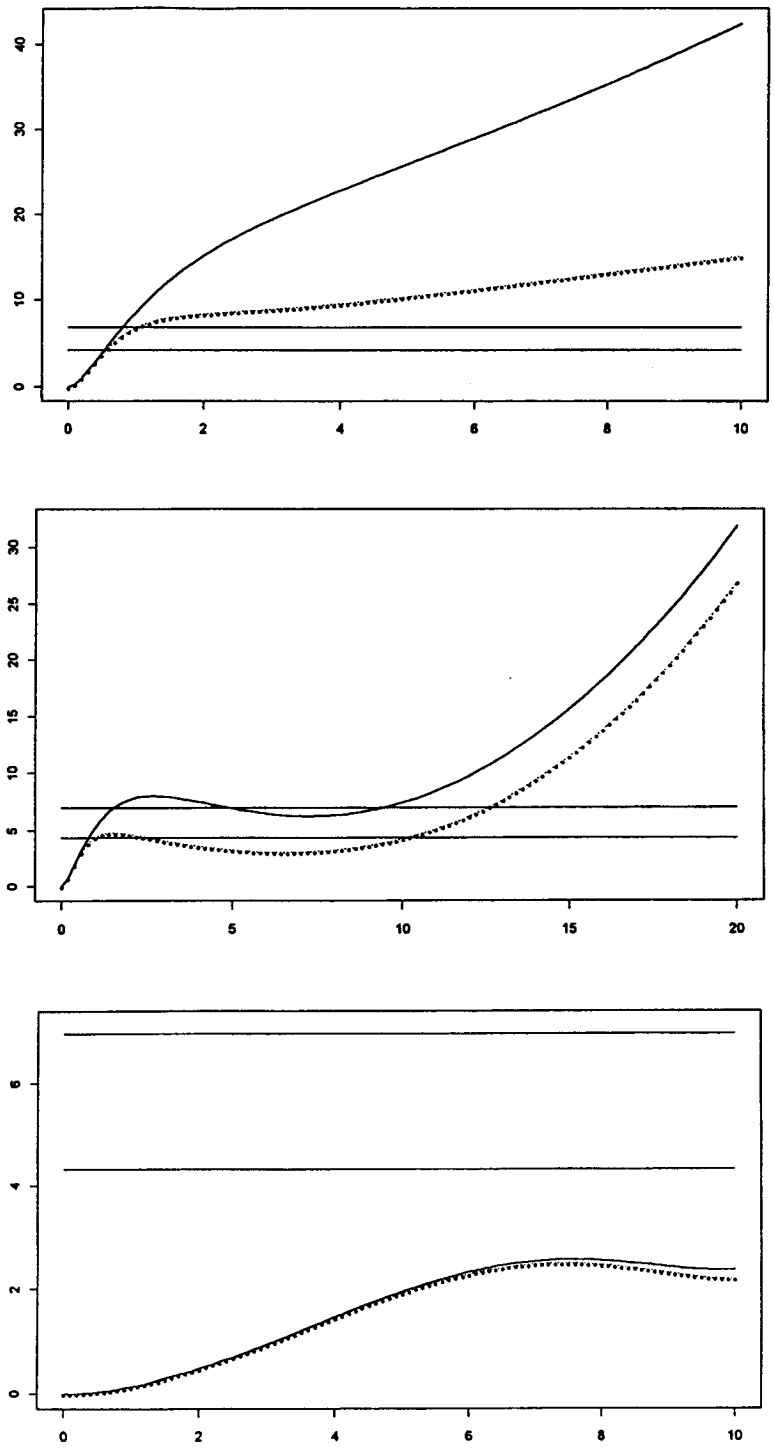

Fig. 1. Graphs of $\lambda(\rho)$ (full lines) and $\tau(\rho)$ (dotted lines). For the angles $\phi = 1.1$, $\phi = 2.2$ and $\phi = 6.25$.

95% likelihood confidence region is correct despite of the intrinsic curvature of the model. The situation may be worse for higher confidence levels.

REFERENCES

Amari, S. I. (1985). *Differential-geometrical Methods in Statistics*, Lecture Notes in Statist., No. 28, Springer, Berlin.

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

Barndorff-Nielsen, O. E. and Cox, D. R. (1979). Edgeworth and saddle-point approximations with statistical applications, *J. Roy. Statist. Soc. Ser. B*, **41**, 279–312.

Bates, D. M. and Watts, D. G. (1980). Relative curvature measures of nonlinearity, *J. Roy. Statist. Soc. Ser. B*, **42**, 1–25.

Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and its Applications*, Wiley, New York.

Clarke, G. P. Y. (1980). Moments of the least-squares estimators in nonlinear regression models, *J. Roy. Statist. Soc. Ser. B*, **42**, 227–237.

Eisenhart, L. P. (1960). *Riemannian Geometry*, Princeton University Press, Princeton.

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

Hougaard, P. (1986). Covariance stabilizing transformations in nonlinear regression, *Scand. J. Statist.*, **13**, 207–210.

Hougaard, P. (1995). Nonlinear regression and curved exponential families. Improvement of the approximation to asymptotic distribution, *Metrika*, **42**, 191–202.

Jensen, J. L. (1995). *Saddlepoint Approximations*, Oxford University Press, New York.

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

Pázman, A. (1990). Almost exact distributions of estimators I and II, *Statistics*, **21**, 9–19 and 21–32.

Pázman, A. (1991). Pivotal variables and confidence regions in flat nonlinear regression models with unknown σ, *Statistics*, **22**, 177–189.

Pázman, A. (1993*a*). Higher dimensional nonlinear regression—A statistical use of the Riemannian curvature tensor, *Statistics*, **25**, 17–25.

Pázman, A. (1993*b*). *Nonlinear Statistical Models*, Kluwer, Dordrecht.

Pázman, A. and Pronzato, L. (1992). Nonlinear experimental design based on the distribution of estimators, *J. Statist. Plann. Inference*, **33**, 385–402.

Seber, G. A. F. and Wild, C. J. (1989). *Nonlinear Regression*, Wiley, New York.

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

Skovgaard, I. M. (1990). On the density of minimum contrast estimators, *Ann. Statist.*, **18**, 779–789.