KYBERNETIKA — VOLUME 30 (1994), NUMBER 2, PAGES 187–198

SECOND-ORDER APPROXIMATION OF THE ENTROPY IN NONLINEAR LEAST-SQUARES ESTIMATION

LUC PRONZATO AND ANDREJ PÁZMAN

Measures of variability of the least-squares estimator $\hat{\theta}$ are essential to assess the quality of the estimation. In nonlinear regression, an accurate approximation of the covariance matrix of $\hat{\theta}$ is difficult to obtain [4]. In this paper, a second-order approximation of the entropy of the distribution of $\hat{\theta}$ is proposed, which is only slightly more complicated than the widely used bias approximation of Box [3]. It is based on the "*flat*" or "*saddle-point approximation*" of the density of $\hat{\theta}$. The neglected terms are of order $\mathcal{O}(\sigma^4)$, while the classical first order approximation neglects terms of order $\mathcal{O}(\sigma^2)$. Various illustrative examples are presented, including the use of the approximate entropy as a criterion for experimental design.

1. INTRODUCTION

Consider a (regular) nonlinear regression model with normal errors

$$\begin{cases} y &= \eta(\bar{\theta}) + \varepsilon\,, \\ \varepsilon &\sim \mathcal{N}(0, \sigma^2 W)\,, \end{cases} \tag{1}$$

where $y \in \mathcal{R}^N$ is the vector of observations, $\bar{\theta}$ the true value of the m-dimensional vector θ of the model parameters, $\theta \in \Theta$ with Θ an open space, and where $\sigma^2 W$ is the variance matrix of ε (the case $W = I$ thus corresponds to independent observations with constant variances). We are interested in the case where $\eta(.)$ is a nonlinear (but smooth) function of θ. In the absence of prior information on θ, one classically uses the maximum likelihood (here generalized least-squares) estimator of θ, given by

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

with $\|a\|_W^2$ denoting $a^{\mathrm{T}} W^{-1} a$. In what follows we assume that to any y corresponds a finite $\hat{\theta}(y)$ (see Example 2 for a discussion).

Measures of the variability of $\hat{\theta}$ are of general importance to evaluate the quality of the estimation, such as the entropy, given by

$$ent(h) := -\int_\Theta \left(\log h(\hat{\theta} \mid \bar{\theta})\right) h(\hat{\theta} \mid \bar{\theta})\, \mathrm{d}\hat{\theta} = -\mathsf{E}_{\bar{\theta}}[\log h(\hat{\theta}(y) \mid \bar{\theta})]\,, \tag{2}$$

where $h(\hat{\theta} \mid \bar{\theta})$ is the probability density of $\hat{\theta}$ when $\bar{\theta}$ is the true value of the parameters, and where the expectation $\mathsf{E}_{\bar{\theta}}$ is taken with respect to y. When the model is linear in θ, $\eta(\theta) = F\theta$, then

$$ent(h) = -\frac{1}{2}\log\det[\sigma^{-2}F^{\mathrm{T}}W^{-1}F] + \frac{m}{2}\log 2\pi + \frac{m}{2}\,, \tag{3}$$

which is closely related to the variance matrix of $\hat{\theta}$, $var(\hat{\theta}) = \sigma^2(F^{\mathrm{T}}W^{-1}F)^{-1}$. However, in the nonlinear situation such a relation does not exist. Moreover, as shown below, the suggested second-order approximation of the entropy of $\hat{\theta}$ is simpler than the second order approximation of $var(\hat{\theta})$ given by Clarke [4], and only slightly more complicated than the widely used bias approximation of Box [3]. One of the reasons why such an approximation for the entropy was not derived before could be the lack of a good approximation of the density of $\hat{\theta}$. Our approach is based on the *"flat"* or *"saddle-point approximation"*.

The approximation of the entropy is derived in Section 2. Various examples are treated in Section 3 to illustrate the validity of this approximation (including one-parameter models, the two-parameter model of Michaelis-Menten and the four-parameter model used by Clarke [4] to illustrate the feasibility of his second-order approximation of moments). In each case we compare the approximated entropy to the standard first-order approximation (given by (3) with $F = \partial\eta(\theta)/\partial\theta|_{\bar{\theta}}$) and to the entropy (2) evaluated by numerical integration or simulation. In linear models, designing a D-optimal experiment corresponds to minimizing the entropy (3). In the nonlinear case, the design of the experiment could thus be based on the approximated entropy, as suggested in Example 4.

2. THE SECOND-ORDER APPROXIMATION

Equation (2) can also be written as

$$ent(h) = -\int_{\mathcal{R}^N}\left(\log h(\hat{\theta}(y) \mid \bar{\theta})\right) f(y - \eta(\bar{\theta}))\,\mathrm{d}y\,,$$

where $f(.)$ is the (normal) probability density of the error vector ε. The second-order approximation is obtained by using the Taylor expansion for $\log h(\hat{\theta}(y) \mid \bar{\theta})$ at the point $y = \eta(\bar{\theta})$,

$$\begin{aligned} ent(h) &= -\int_{\mathcal{R}^N}\left(\log h(\bar{\theta} \mid \bar{\theta}) + \left.\frac{\partial \log h(\hat{\theta}(y) \mid \bar{\theta})}{\partial y^{\mathrm{T}}}\right|_{\eta(\bar{\theta})}\varepsilon\right. \\ &\quad \left. + \frac{1}{2}\varepsilon^{\mathrm{T}}\left.\frac{\partial^2 \log h(\hat{\theta}(y) \mid \bar{\theta})}{\partial y \partial y^{\mathrm{T}}}\right|_{\eta(\bar{\theta})}\varepsilon + \mathcal{O}(\|\varepsilon\|^3)\right) f(\varepsilon)\,\mathrm{d}\varepsilon \\ &= -\log h(\bar{\theta} \mid \bar{\theta}) - \frac{\sigma^2}{2}\left.\frac{\partial^2 \log h(\hat{\theta}(y) \mid \bar{\theta})}{\partial y_i \partial y_j}\right|_{\eta(\bar{\theta})} W_{ij} + er(h)\,, \end{aligned} \tag{4}$$

where $er(h)$ denotes the error of the second-order approximation of the entropy. In Appendix B, we show that it is of order of magnitude $\mathcal{O}(\sigma^4)$ provided that the integral (2) exists and that $h(\hat{\theta} \mid \bar{\theta})$ has continuous 4-th order derivatives.

Note that in (4) and in the sequel we take the sum over subscripts that appear twice in the same term.

The second-order derivative can be developped to obtain

$$\begin{aligned}\left.\frac{\partial^2 \log h(\hat{\theta}(y) \mid \bar{\theta})}{\partial y_i \partial y_j}\right|_{\eta(\bar{\theta})} W_{ij} &= \left.\frac{\partial^2 \log h(\theta \mid \bar{\theta})}{\partial \theta_k \partial \theta_l}\right|_{\bar{\theta}} \left(\left.\frac{\partial \hat{\theta}_k(y)}{\partial y_i}\right|_{\eta(\bar{\theta})} W_{ij} \left.\frac{\partial \hat{\theta}_l(y)}{\partial y_j}\right|_{\eta(\bar{\theta})} \right) \\ &+ \left.\frac{\partial \log h(\theta \mid \bar{\theta})}{\partial \theta_k}\right|_{\bar{\theta}} \left(\left.\frac{\partial^2 \hat{\theta}_k(y)}{\partial y_i \partial y_j}\right|_{\eta(\bar{\theta})} W_{ij} \right) .\end{aligned}$$

The computation of the derivatives of $\hat{\theta}(y)$ has been considered in papers dealing with the second-order approximation of moments (cf. [3, 4]). A more straighforward method is presented in Appendix A. Substituting for the second-order derivative of $h(\hat{\theta}(y) \mid \bar{\theta})$ from the previous expression into (4), we then get from Appendix A

$$\begin{aligned}ent(h) &= -\log h(\bar{\theta} \mid \bar{\theta}) - \frac{\sigma^2}{2} \left.\frac{\partial^2 \log h(\theta \mid \bar{\theta})}{\partial \theta_k \partial \theta_l}\right|_{\bar{\theta}} M_{kl}^{-1} \\ &+ \frac{\sigma^2}{2} \left.\frac{\partial \log h(\theta \mid \bar{\theta})}{\partial \theta_k}\right|_{\bar{\theta}} M_{ka}^{-1} Z_{bc}^{a} M_{bc}^{-1} + er(h) , \qquad (5)\end{aligned}$$

with

$$\begin{aligned}M_{ij} &:= M_{ij}(\bar{\theta}) := \left.\frac{\partial \eta^{\mathrm{T}}(\theta)}{\partial \theta_i}\right|_{\bar{\theta}} W^{-1} \left.\frac{\partial \eta(\theta)}{\partial \theta_j}\right|_{\bar{\theta}} , \\ Z_{ij}^{a} &:= \left.\frac{\partial \eta^{\mathrm{T}}(\theta)}{\partial \theta_a}\right|_{\bar{\theta}} W^{-1} \left.\frac{\partial^2 \eta(\theta)}{\partial \theta_i \partial \theta_j}\right|_{\bar{\theta}} , \qquad (6)\end{aligned}$$

(the first term is the Fisher information matrix for $\sigma = 1$, and the second one is the affine connection in the sense of Amari [1]).

By substituting the classical (asymptotic) normal approximation for $h(\hat{\theta} \mid \bar{\theta})$ into (5) one obtains the first-order approximation of the entropy which was mentioned in the introduction.

A much better approximation of $h(\hat{\theta} \mid \bar{\theta})$ is given by the *flat* or *saddle-point approximation* [6, 5]

$$q(\hat{\theta} \mid \bar{\theta}) = \frac{\det Q(\hat{\theta}, \bar{\theta})}{(2\pi)^{m/2} \sigma^m \det^{1/2} M(\hat{\theta})} \exp\left[-\frac{1}{2\sigma^2} \| P(\hat{\theta})(\eta(\hat{\theta}) - \eta(\bar{\theta})) \|_W^2 \right] , \qquad (7)$$

where

$$\begin{aligned}P_{ij}(\theta) &:= \frac{\partial \eta_i(\theta)}{\partial \theta_a} M_{ab}(\theta)^{-1} \frac{\partial \eta_c(\theta)}{\partial \theta_b} W_{cj}^{-1} , \\ Q_{ij}(\theta, \bar{\theta}) &:= M_{ij}(\theta) + (\eta_a(\theta) - \eta_a(\bar{\theta})) W_{ab}^{-1} (I - P(\theta))_{bc} \frac{\partial^2 \eta_c(\theta)}{\partial \theta_i \partial \theta_j} .\end{aligned}$$

The saddle-point method is known to yield accurate approximations of densities in general (cf. e.g. [8]). Moreover, the density $q(\hat{\theta} \mid \bar{\theta})$ is exact in any model with a zero intrinsic curvature (in particular when the number of design points is equal to m, see Example 4). It is "almost exact" in models with a zero Riemannian curvature (see (10)), including all one-parameter nonlinear models. On the other hand, this approximate density can be used only if the intrinsic curvature is not too large when compared to σ and if there is no overlapping of the model (there are no restrictions on the magnitude of the parameter-effect curvature). However, in such cases least-squares estimation will fail to give a reasonable answer (see (Pázman, 1990) for a discussion on the properties of $q(\hat{\theta} \mid \bar{\theta})$).

Proposition 1. The entropy of the density $q(\hat{\theta} \mid \bar{\theta})$ is equal to

$$ent(q) = ent_2 + er(q)\,,$$

where the error term $er(q)$ is given in Appendix B and where ent_2 is the second-order approximation

$$ent_2 = ent_1 + \sigma^2 M_{ij}^{-1}[M_{ab}^{-1}(R_{ajbi} + U_{aij}^{b}) - \Gamma_{ai}^{d}\Gamma_{dj}^{a} - \Gamma_{ac}^{a}\Gamma_{ij}^{c}]\,. \tag{8}$$

Here ent_1 denotes the first-order approximation of the entropy,

$$ent_1 = -\frac{1}{2}\log\det[\sigma^{-2}M] + \frac{m}{2}\log(2\pi) + \frac{m}{2}\,, \tag{9}$$

R_{ajbi} is the component of the Riemannian curvature tensor at $\bar{\theta}$ [1, 7]

$$R_{ajbi}(\theta) := \frac{\partial^2\eta^{\mathrm{T}}(\theta)}{\partial\theta_a\partial\theta_b}W^{-1}(I-P(\theta))\frac{\partial^2\eta(\theta)}{\partial\theta_j\partial\theta_i} - \frac{\partial^2\eta^{\mathrm{T}}(\theta)}{\partial\theta_a\partial\theta_i}W^{-1}(I-P(\theta))\frac{\partial^2\eta(\theta)}{\partial\theta_j\partial\theta_b}\,, \tag{10}$$

Γ_{ai}^{c} is the affine connection in regression models [1]

$$\Gamma_{ai}^{d} := Z_{ai}^{c}M_{cd}^{-1}\,,$$

and U_{aij}^{b} is defined by

$$U_{aij}^{b} := \left.\frac{\partial^3\eta_c(\theta)}{\partial\theta_a\partial\theta_i\partial\theta_j}\right|_{\bar{\theta}} W_{cd}^{-1} \left.\frac{\partial\eta_d(\theta)}{\partial\theta_b}\right|_{\bar{\theta}}\,. \tag{11}$$

Note that $R_{ijhk} = 0$ in flat models (which include all one-parameter nonlinear models and also the Michaelis-Menten model of Example 4).

Proof. From the expression (7) for the density $q(\hat{\theta} \mid \bar{\theta})$, we have

$$\log q(\hat{\theta} \mid \bar{\theta}) = \log\det Q(\hat{\theta},\bar{\theta}) - \frac{1}{2}\log\det M(\hat{\theta}) - \frac{1}{2\sigma^2}\|P(\hat{\theta})(\eta(\hat{\theta}) - \eta(\bar{\theta}))\|_W^2 - \frac{m}{2}\log 2\pi - \frac{m}{2}\log\sigma^2\,,$$

which can be derived term by term. We have

$$\begin{aligned} \frac{\partial \log\det M(\theta)}{\partial\theta_i} &= M_{ab}^{-1}(\theta)\frac{\partial M_{ab}(\theta)}{\partial\theta_i} = 2Z_{ai}^b(\theta)M_{ab}^{-1}(\theta)\,, \\ \frac{\partial M_{ab}^{-1}(\theta)}{\partial\theta_i} &= -M_{ac}^{-1}(\theta)\frac{\partial M_{cd}(\theta)}{\partial\theta_i}M_{db}^{-1}(\theta)\,. \end{aligned} \tag{12}$$

This gives

$$\left.\frac{\partial^2 \log\det M(\theta)}{\partial\theta_i\partial\theta_j}\right|_{\bar\theta} = 2(X_{ai}^{bj} + U_{aij}^b)M_{ab}^{-1} - 2Z_{ai}^b M_{ac}^{-1}(Z_{cj}^d + Z_{dj}^c)M_{db}^{-1}\,, \tag{13}$$

with

$$X_{ai}^{bj} := \left.\frac{\partial^2\eta_a(\theta)}{\partial\theta_a\partial\theta_i}\right|_{\bar\theta} W_{ab}^{-1} \left.\frac{\partial^2\eta_b(\theta)}{\partial\theta_b\partial\theta_j}\right|_{\bar\theta}\,,$$

and U and Z respectively given by (11) and (6). Since

$$\frac{\partial\eta_c(\theta)}{\partial\theta_i}W_{cd}^{-1}(I - P(\theta))_{de} = 0\,, \tag{14}$$

we have

$$\begin{aligned} \frac{\partial Q_{ab}(\theta,\bar\theta)}{\partial\theta_i} &= \frac{\partial M_{ab}(\theta)}{\partial\theta_i} - (\eta_c(\theta) - \eta_c(\bar\theta))W_{cd}^{-1}\frac{\partial P_{de}(\theta)}{\partial\theta_i}\frac{\partial^2\eta_e(\theta)}{\partial\theta_a\partial\theta_b} \\ &\quad + (\eta_c(\theta) - \eta_c(\bar\theta))W_{cd}^{-1}(I_{de} - P_{de}(\theta))\frac{\partial^3\eta_e(\theta)}{\partial\theta_a\partial\theta_b\partial\theta_i}\,. \end{aligned}$$

From

$$\frac{\partial \log\det Q(\theta,\bar\theta)}{\partial\theta_i} = Q_{ab}^{-1}(\theta,\bar\theta)\frac{\partial Q_{ab}(\theta,\bar\theta)}{\partial\theta_i}\,, \tag{15}$$

we have

$$\begin{aligned} &\left.\frac{\partial^2 \log\det Q(\theta,\bar\theta)}{\partial\theta_i\partial\theta_j}\right|_{\bar\theta} \\ &= \left.\frac{\partial^2 \log\det M(\theta)}{\partial\theta_i\partial\theta_j}\right|_{\bar\theta} - \left.\left[\frac{\partial\eta_c(\theta)}{\partial\theta_j}W_{cd}^{-1}\frac{\partial P_{de}(\theta)}{\partial\theta_i}\frac{\partial^2\eta_e(\theta)}{\partial\theta_a\partial\theta_b}M_{ab}^{-1}\right]\right|_{\bar\theta}\,. \end{aligned} \tag{16}$$

By straightforward calculations, we then obtain

$$\left.\left[\frac{\partial\eta_c(\theta)}{\partial\theta_j}W_{cd}^{-1}\frac{\partial P_{de}(\theta)}{\partial\theta_i}\frac{\partial^2\eta_e(\theta)}{\partial\theta_a\partial\theta_b}\right]\right|_{\bar\theta} = -Z_{ij}^c M_{cd}^{-1} Z_{ab}^d + X_{ij}^{ab}\,.$$

Further, using (14) we obtain

$$\left.\frac{\partial\left\|P(\eta(\theta) - \eta(\bar\theta))\right\|_W^2}{\partial\theta_i}\right|_{\bar\theta} = 0\,, \tag{17}$$

$$\left.\frac{\partial^2\|P(\eta(\theta) - \eta(\bar\theta))\|_W^2}{\partial\theta_i\partial\theta_j}\right|_{\bar\theta} = 2M_{ij}\,. \tag{18}$$

Taking into account the results in (12–18) we obtain the required derivatives of $\log q$,

$$\begin{aligned}
\left.\frac{\partial \log q(\theta \mid \bar\theta)}{\partial \theta_i}\right|_{\bar\theta} &= Z^b_{ai} M^{-1}_{ab} \,, \qquad (19)\\
\left.\frac{\partial^2 \log q(\theta \mid \bar\theta)}{\partial \theta_i \partial \theta_j}\right|_{\bar\theta} &= \frac{1}{2}\left.\frac{\partial^2 \log det M(\theta)}{\partial \theta_i \partial \theta_j}\right|_{\bar\theta} + Z^c_{ij} M^{-1}_{cd} Z^d_{ab} M^{-1}_{ab} - X^{ab}_{ij} M^{-1}_{ab} - \frac{1}{\sigma^2} M_{ij} \\
&= X^{bj}_{ai} M^{-1}_{ab} - X^{ab}_{ij} M^{-1}_{ab} - Z^b_{ai} M^{-1}_{ac}(Z^d_{cj} + Z^c_{dj}) M^{-1}_{db} \\
&\quad + Z^c_{ij} M^{-1}_{cd} Z^d_{ab} M^{-1}_{ab} + U^b_{aij} M^{-1}_{ab} - \frac{1}{\sigma^2} M_{ij} \,.
\end{aligned}$$

From the definition of the projector $P(\theta)$, it follows that

$$Z^b_{ai} M^{-1}_{db} Z^d_{cj} = \left.\frac{\partial^2 \eta^{\mathrm{T}}(\theta)}{\partial \theta_a \partial \theta_i}\right|_{\bar\theta} W^{-1} P \left.\frac{\partial^2 \eta(\theta)}{\partial \theta_c \partial \theta_j}\right|_{\bar\theta} .$$

Hence we have

$$X^{bj}_{ai} M^{-1}_{ab} - Z^b_{ai} M^{-1}_{ac} Z^d_{cj} M^{-1}_{db} = \left.\frac{\partial^2 \eta^{\mathrm{T}}(\theta)}{\partial \theta_a \partial \theta_i}\right|_{\bar\theta} W^{-1}(I - P) \left.\frac{\partial^2 \eta(\theta)}{\partial \theta_b \partial \theta_j}\right|_{\bar\theta} M^{-1}_{ab} \,,$$

and similarly

$$-X^{ab}_{ij} M^{-1}_{ab} + Z^c_{ij} M^{-1}_{cd} Z^d_{ab} M^{-1}_{ab} = -\left.\frac{\partial^2 \eta^{\mathrm{T}}(\theta)}{\partial \theta_a \partial \theta_b}\right|_{\bar\theta} W^{-1}(I - P) \left.\frac{\partial^2 \eta(\theta)}{\partial \theta_i \partial \theta_j}\right|_{\bar\theta} M^{-1}_{ab} \,.$$

We thus have from the definition of the Riemanian curvature tensor (10)

$$\left.\frac{\partial^2 \log q(\theta \mid \bar\theta)}{\partial \theta_i \partial \theta_j}\right|_{\bar\theta} = R_{ajbi} M^{-1}_{ab} - Z^c_{ai} M^{-1}_{ab} Z^b_{dj} M^{-1}_{dc} + U^b_{aij} M^{-1}_{ab} - \frac{1}{\sigma^2} M_{ij} \,, \qquad (20)$$

which, together with (5), completes the proof. □

3. EXAMPLES

Through all this section we consider the case of independent observations.

Example 1. Consider the following one-parameter model,

$$\eta(\theta, x) = \theta + \theta^2 + x\theta^3 \,, \ \ \theta \in \mathcal{R} \,,$$

with the two design points $x_1 = 1, x_2 = 1.2$ and with $\bar\theta = 1$. We compare the value ent_2 (8) of the second-order approximation of the entropy with its first-order approximation ent_1 (9) and with the value ent_q obtained by the numerical evaluation of the integral $-\int_{\mathcal{R}} \log q(\hat\theta \mid \bar\theta) q(\hat\theta \mid \bar\theta)\, \mathrm{d}\hat\theta$. Figure 1 presents the differences $\mid ent_q -$

ent_1 | and | $ent_q - ent_2$ | as functions of σ^2. This clearly illustrates that | $ent_q - ent_1$ | is of order $\mathcal{O}(\sigma^2)$, while | $ent_q - ent_2$ | is of order $\mathcal{O}(\sigma^4)$.

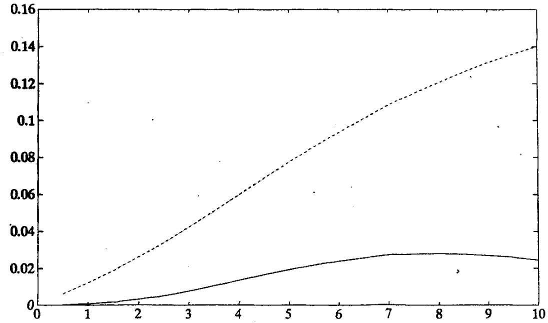

Fig. 1. | $ent_q - ent_1$ | (dashed line) and | $ent_q - ent_2$ | (full line) as functions of σ^2 (Example 1).

Example 2. Consider the model [3]

$$\eta(\theta, x) = \exp(-\theta x)\,,$$

with fixed design $\{x_1, \ldots, x_p\}$. A specific feature of this example is the fact that when θ tends to ∞ the response remains finite, so that the expectation surface

$$\mathcal{E} = \{(\eta(\theta, x_1), \ldots, \eta(\theta, x_p))^{\mathrm{T}} \mid \theta \in \mathcal{R}\}$$

is bounded. As a consequence, some samples y give estimates with extremely large values (or even $\hat{\theta}(y)$ does not exist). For small values of σ^2, the probability of such an event is negligible, as it is the case in the numerical example considered by Box [3]. However, this is not the case for larger values of σ^2, and it raises some theoretical problems. Note for instance that the bias is infinite for any $\sigma > 0$, as it is mentioned by J .D. Sargan in the discussion of [3]. The same difficulty is met with the entropy (2). Theoretically it is always infinite, although, in practice, if we neglect the probability for $\hat{\theta}(y)$ to be infinite, the approximation suggested in this paper can be used, with the same limitations as for the approximation of the bias suggested by Box [3].

To be able to deal with such problems due to a bounded expectation surface, it seems reasonable to reconsider the distribution of y in application. For instance, here, as suggested in the discussion by J. D. Sargan, a lognormal distribution for y is more realistic from the experimental point of view (which yields a linear model with normal errors after a transformation of variables).

Example 3. Consider the Richard's function used in [4],

$$\eta(\theta, x) = \theta_4 - \theta_3 \log \left[1 + \exp(-\frac{\theta_2}{\theta_3} - \frac{\theta_1 x}{\theta_3}) \right] ,$$

with the same eleven design points as in this paper, $x_1 = -2.15, x_2 = -1.5, x_3 = -0.85, x_4 = -0.08, x_5 = 0.52, x_6 = 1.1, x_7 = 2.28, x_8 = 3.23, x_9 = 4, x_{10} = 4.65, x_{11} = 5$. The value $\bar{\theta}$ is taken equal to the estimate $\hat{\theta}$ of [4],

$$\bar{\theta} = (1.6415, -2.3920, 1.7678, 8.5540)^{\mathrm{T}} .$$

We compare the quadratic approximation ent_2 (8) of the entropy with its linear approximation ent_1 (9), and with the value ent_q obtained by numerical simulation, i.e. by calculating the mean

$$m_n = -\frac{1}{n} \sum_{i=1}^{n} \log q(\hat{\theta}(y^{(i)}) \mid \bar{\theta}) ,$$

with $n = 10^4$, where the $y^{(i)}$'s are generated according to (1). The relative errors $\mid ent_1 - m_n \mid / \mid m_n \mid$, $\mid ent_2 - m_n \mid / \mid m_n \mid$ and the observed (relative) standard-deviation of m_n, denoted by s_n/m_n, are given in Table 1 for two different values of σ^2.

Table 1. Relative errors of the first and second-order approximations (Example 3).

σ^2	$\mid ent_1 - m_n \mid / \mid m_n \mid$	$\mid ent_2 - m_n \mid / \mid m_n \mid$	s_n/m_n
$3.3993 \times 10^{-3(*)}$	6.9×10^{-3}	1.8×10^{-3}	2.5×10^{-3}
0.02	0.10	2.3×10^{-2}	8.5×10^{-3}

$(*)$: estimated value of σ^2 in [4]

For the same value of σ^2 as the one used in [4], the relative error of ent_1 is only of 0.69 % (resp. 0.18 % for ent_2), while the relative error on the first-order approximation of the matrix $var(\hat{\theta})$ taken from [4] reaches 40 % for some of its terms. This can be explained by the fact that the entropy summarizes the variabilities on the different components of $\hat{\theta}$ which may compensate each other. Note, however, that the error in ent_2 is about four times smaller than the error in ent_1. Approximately the same improvement is indicated in [4], when comparing the first-order approximation of $var(\hat{\theta})$ to its second-order approximation.

Example 4. Designing a D-optimal experiment in nonlinear regression corresponds to minimizing the first-order approximation ent_1 (9) of the entropy. It thus seems reasonable to use the second-order approximation ent_2 (8) as a criterion for experimental design. We simply present here a numerical example. We consider the Michaelis-Menten model response,

$$\eta(\theta, x) = \frac{\theta_1 x}{\theta_2 + x} ,$$

with $\sigma^2 = 1.5 \times 10^{-4}$, $\bar{\theta}$ equal to the value $\hat{\theta}$ of [2],

$$\bar{\theta} = (0.10579, 1.7007)^{\mathrm{T}},$$

and six design points $x_i, i = 1, \ldots, 6$ in [0, 2]. The results are given in Table 2, where $k \otimes x$ denotes k replications of the design point x. We first compute the D-optimal experiment (independent of the value of σ^2), which is obtained for three replications on two design points (line 1 of Table 2). Even when three replications on two design points are imposed, the optimal design for the criterion ent_2 is slightly different (line 2 of Table 2). When this constraint is cancelled, much better design can be obtained, as indicated by the third line of Table 2. Note that each design in the table only possesses two distinct support points, which corresponds to a zero intrinsic curvature, so that the density (7) used to construct ent_2 is exact.

Table 2. Comparison between various optimal designs (Example 4).

design	ent_1	ent_2
$\{3 \otimes 0.6297, 3 \otimes 2\}$	−1.056	−0.3583
$\{3 \otimes 0.6829, 3 \otimes 2\}$	−1.052	−0.3674
$\{4 \otimes 0.6538, 2 \otimes 2\}^{(*)}$	−0.996	−0.4315

$^{(*)}$: locally optimal for ent_2

APPENDIX A

Derivatives of $\hat{\theta}(y)$

The function $\hat{\theta}(y)$ is defined implicitly by the normal equation

$$\frac{\partial}{\partial \theta}\left[\frac{1}{2}\|y - \eta(\theta)\|_W^2\right] = 0,$$

which can be written

$$(\eta(\theta) - y)_a W_{ab}^{-1} \frac{\partial \eta_b(\theta)}{\partial \theta_i} = 0\,; \; i = 1, \ldots, m. \tag{A1}$$

Denote the left-hand side of (A1) by $F_i(\theta, y)$. From the implicit function theorem we then have

$$\frac{\partial \hat{\theta}_i(y)}{\partial y_a} = -\left[\left(\frac{\partial F_i(\theta, y)}{\partial \theta_j}\right)^{-1} \frac{\partial F_j(\theta, y)}{\partial y_a}\right]_{\theta = \hat{\theta}(y)}, \tag{A2}$$

where

$$\begin{aligned} \frac{\partial F_i(\theta, y)}{\partial \theta_j} &= M_{ij}(\theta) + (\eta_a(\theta) - y_a) W_{ab}^{-1} \frac{\partial^2 \eta_b(\theta)}{\partial \theta_i \partial \theta_j}, \\ \frac{\partial F_j(\theta, y)}{\partial y_a} &= -W_{ab}^{-1} \frac{\partial \eta_b(\theta)}{\partial \theta_j}. \end{aligned} \tag{A3}$$

Consequently, since $\hat{\theta}(\eta(\bar{\theta})) = \bar{\theta}$, one has

$$\left.\frac{\partial \hat{\theta}_i(y)}{\partial y_a}\right|_{\eta(\bar{\theta})} = M_{ij}^{-1} \left.\frac{\partial \eta_b(\theta)}{\partial \theta_j}\right|_{\bar{\theta}} W_{ab}^{-1},$$

and

$$\left.\frac{\partial \hat{\theta}_i(y)}{\partial y_a}\right|_{\eta(\bar{\theta})} W_{ab} \left.\frac{\partial \hat{\theta}_j(y)}{\partial y_b}\right|_{\eta(\bar{\theta})} = M_{ij}^{-1}.$$

Taking the derivative in (A2), we obtain

$$\begin{aligned} \left.\frac{\partial^2 \hat{\theta}_i(y)}{\partial y_a \partial y_b}\right|_{\eta(\bar{\theta})} W_{ab} &= -M_{ij}^{-1} \left.\frac{dG_{jk}(\theta, y)}{dy_b}\right|_{\theta=\hat{\theta}(y), y=\eta(\bar{\theta})} M_{kl}^{-1} \left.\frac{\partial \eta_b(\theta)}{\partial \theta_l}\right|_{\bar{\theta}} \\ &\quad + M_{ij}^{-1} \left.\frac{\partial^2 \eta_b(\theta)}{\partial \theta_j \partial \theta_k}\right|_{\bar{\theta}} \left.\frac{\partial \hat{\theta}_k(y)}{\partial y_b}\right|_{\eta(\bar{\theta})}, \end{aligned} \tag{A4}$$

where $G_{ij}(\theta, y)$ denotes the right-hand side of (A3). At $\theta = \bar{\theta}$, $y = \eta(\bar{\theta})$ we have

$$\begin{aligned} \left.\frac{dG_{ik}(\theta, y)}{dy_b}\right|_{\theta=\hat{\theta}(y), y=\eta(\bar{\theta})} &= \left.\frac{\partial G_{ik}(\theta, \eta(\bar{\theta}))}{\partial \theta_j}\right|_{\bar{\theta}} \left.\frac{\partial \hat{\theta}_j(y)}{\partial y_b}\right|_{\eta(\bar{\theta})} + \left.\frac{\partial G_{ik}(\bar{\theta}, y)}{\partial y_b}\right|_{\eta(\bar{\theta})} \\ &= (Z_{kj}^i + Z_{ij}^k) M_{jm}^{-1} \left.\frac{\partial \eta_c(\theta)}{\partial \theta_m}\right|_{\bar{\theta}} W_{cb}^{-1} + Z_{ik}^j M_{jm}^{-1} \left.\frac{\partial \eta_c(\theta)}{\partial \theta_m}\right|_{\bar{\theta}} W_{cb}^{-1} - W_{bc}^{-1} \left.\frac{\partial^2 \eta_c(\theta)}{\partial \theta_i \partial \theta_k}\right|_{\bar{\theta}}. \end{aligned}$$

Finally, after some simple algebraic manipulations we obtain from (A4)

$$\left.\frac{\partial^2 \hat{\theta}_i(y)}{\partial y_a \partial y_b}\right|_{\eta(\bar{\theta})} W_{ab} = -M_{ij}^{-1} Z_{kl}^j M_{kl}^{-1}.$$

APPENDIX B

Error of approximation

If the density $h(\hat{\theta} \mid \bar{\theta})$ has continuous 4-th order derivatives with respect to $\hat{\theta}$ and is such that the integral (2) exists, then for any (small) $\delta > 0$ the error term in (4) depends of σ according to the inequality

$$| \, er(h) \, | \leq \left[1 + | \log h(\bar{\theta} \mid \bar{\theta}) | + \frac{1}{2} \left| \sum_{i,j} \left.\frac{\partial^2 \log h(\hat{\theta}(y) \mid \bar{\theta})}{\partial y_i \partial y_j}\right|_{\eta(\bar{\theta})} \right| \right] \delta + \mathcal{O}(\sigma^4).$$

Proof. In order to simplify the notations, we denote

$$\phi(\varepsilon) := h(\hat{\theta}(\varepsilon + \eta(\bar{\theta})) \mid \bar{\theta}).$$

We also define

$$\mathcal{S}_r := \{z \in \mathcal{R}^N : \|z\|_W \le r\}\,.$$

For any $\delta > 0$, (not depending on σ), there exists a positive number $r(\sigma,\delta)$ which is large enough to ensure that

$$\int_{\mathcal{R}^N - \mathcal{S}_{r(\sigma,\delta)}} f(\varepsilon)\,\mathrm{d}\varepsilon \;<\; \delta\,, \tag{B1}$$

$$\left|\int_{\mathcal{R}^N - \mathcal{S}_{r(\sigma,\delta)}} (\log\phi(\varepsilon)) f(\varepsilon)\,\mathrm{d}\varepsilon\right| \;<\; \delta\,, \tag{B2}$$

$$\left|\int_{\mathcal{R}^N - \mathcal{S}_{r(\sigma,\delta)}} \varepsilon_i\varepsilon_j f(\varepsilon)\,\mathrm{d}\varepsilon\right| \;<\; \delta\,,\; i,j = 1,\ldots,m\,. \tag{B3}$$

From the Taylor formula for $\log\phi(\varepsilon)$, we obtain

$$\begin{aligned}\int_{\mathcal{S}_{r(\sigma,\delta)}} (\log\phi(\varepsilon)) f(\varepsilon)\,\mathrm{d}\varepsilon &= \log\phi(0)\int_{\mathcal{S}_{r(\sigma,\delta)}} f(\varepsilon)\,\mathrm{d}\varepsilon \\ &+\frac{1}{2!}\left.\frac{\partial^2\log\phi(\varepsilon)}{\partial\varepsilon_i\partial\varepsilon_j}\right|_0 \int_{\mathcal{S}_{r(\sigma,\delta)}} \varepsilon_i\varepsilon_j f(\varepsilon)\,\mathrm{d}\varepsilon \\ &+\frac{1}{4!}\int_{\mathcal{S}_{r(\sigma,\delta)}} \left.\frac{\partial^4\log\phi(\lambda)}{\partial\lambda_i\partial\lambda_j\partial\lambda_k\partial\lambda_l}\right|_{\lambda=\psi(\varepsilon)} \varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_l f(\varepsilon)\,\mathrm{d}\varepsilon\,,\end{aligned} \tag{B4}$$

where $\psi(\varepsilon)$ is such that $\psi(\varepsilon) \in \mathcal{S}_{r(\sigma,\delta)}$ for any $\varepsilon \in \mathcal{S}_{r(\sigma,\delta)}$. Since $\frac{\partial^4\log\phi(\lambda)}{\partial\lambda_i\partial\lambda_j\partial\lambda_k\partial\lambda_l}$ is continuous, it is bounded on the compact set $\mathcal{S}_{r(1,\delta)}$. The densities $f(\varepsilon)$ and $h(\hat\theta \mid \bar\theta)$, both functions of σ, are more concentrated when σ decreases. We can therefore take $\mathcal{S}_{r(\sigma,\delta)} \subset \mathcal{S}_{r(1,\delta)}$ for $\sigma < 1$, and the same bound can be used for $\frac{\partial^4\log\phi(\lambda)}{\partial\lambda_i\partial\lambda_j\partial\lambda_k\partial\lambda_l}$ on $\mathcal{S}_{r(\sigma,\delta)}$ and $\mathcal{S}_{r(1,\delta)}$. Consequently, the last term in (B4) is of the order of magnitude $\mathcal{O}(\sigma^4)$. From (B1–B4) we obtain

$$\begin{aligned}&\left| ent(h) - \log\phi(0) - \frac{\sigma^2}{2}\left.\frac{\partial^2\log\phi(\varepsilon)}{\partial\varepsilon_i\partial\varepsilon_j}\right|_0 W_{ij}\right| \;\le \\ &\left[1 + |\log\phi(0)| + \frac{1}{2}\left|\sum_{i,j}\left.\frac{\partial^2\log\phi(\varepsilon)}{\partial\varepsilon_i\partial\varepsilon_j}\right|_0\right|\right]\delta \;+\; \mathcal{O}(\sigma^4)\,,\end{aligned}$$

for any fixed δ. □

(Received January 4, 1993.)

REFERENCES

[1] S. Amari: Differential-Geometrical Methods in Statistics. Springer, Berlin 1985.
[2] D. Bates and D. Watts: Relative curvature measures of nonlinearity. J. Roy. Statist. Soc. Ser. B *42* (1980), 1–25.

[3] M. Box: Bias in nonlinear estimation. J. Roy. Statist. Soc. Ser. B *33* (1971), 171–201.
[4] G. Clarke: Moments of the least-squares estimators in a non-linear regression model. J. Roy. Statist. Soc. Ser. B *42* (1980), 227–237.
[5] P. Hougaard: Saddlepoint approximations for curved exponential families. Statist. Probab. Lett. *3* (1985), 161–166.
[6] A. Pázman: Probability distribution of the multivariate nonlinear least-squares estimates. Kybernetika *20* (1984), 209–230.
[7] A. Pázman: Small-sample distributional properties of nonlinear regression estimators (a geometric approach) (with discussion). Statistics *21* (1990), 3, 323–367.
[8] N. Reid: Saddlepoint methods and statistical inference. Statist. Sci. *3* (1988), 213–238.

Luc Pronzato, Laboratoire I3S, CNRS-URA 1376, Sophia Antipolis, 06560 Valbonne. France. On leave from Laboratoire des Signaux et Systèmes, CNRS-ESE, 91192 Gif-sur-Yvette. France.

Andrej Pázman, Department of Probability and Statistics, Faculty of Mathematics and Physics, Commenius University, Mlynská dolina 842 15 Bratislava. Slovakia.