A SUFFICIENT STATISTIC AND A NOSTANDARD LINEARIZATION IN NONLINEAR REGRESSION

ANDREJ PÁZMAN

In a nonlinear model $\mathbf{y} = \boldsymbol{\eta}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}$ a standard linearization consists in linearizing $\boldsymbol{\eta}(\boldsymbol{\theta})$ at a point $\boldsymbol{\theta}^*$, and in computing the M. L. estimate $\boldsymbol{\tau}(\mathbf{y}, \boldsymbol{\theta}^*)$ in the linearized model. We propose to take $\boldsymbol{\tau}(\mathbf{y}) := (\boldsymbol{\tau}(\mathbf{y}, \boldsymbol{\theta}^1), \dots, \boldsymbol{\tau}(\mathbf{y}, \boldsymbol{\theta}^k))^{\mathrm{T}}$ for some $\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^k$ (= the sufficient statistic), linearize each $\boldsymbol{\tau}(\mathbf{y}, \boldsymbol{\theta}^i)$ separately, and then to compute the M. L. estimate $\tilde{\boldsymbol{\theta}}(\mathbf{y})$. The variable $\tilde{\boldsymbol{\theta}}(\mathbf{y})$ has a smaller variance than $\boldsymbol{\tau}(\mathbf{y}, \boldsymbol{\theta}^i)$, and a comparable bias. Further, $\tilde{\boldsymbol{\theta}}(\mathbf{y})$ can be used to approximate the posterior density in a Bayesian approach.

The construction of the sufficient statistic has a geometrical background. Possible consequences for nonlinear experimental design are mentioned.

1. INTRODUCTION AND THE GEOMETRICAL BACKGROUND

Let us consider the nonlinear regression model with normal errors

$$\mathbf{y} = \boldsymbol{\eta}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}; \quad (\boldsymbol{\theta} \in \Theta) \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{1}$$

under standard regularity assumptions: the parameter space Θ is an open subset of $\mathbb{R}^m$, the variance matrix $\boldsymbol{\Sigma}$ is regular, the regression mapping $\boldsymbol{\eta}$: $\Theta \mapsto \mathbb{R}^N$ $(N > m)$ has continuous second order derivatives on Θ, and the vectors $\partial\boldsymbol{\eta}(\boldsymbol{\theta})/\partial\theta_1, \dots \partial\boldsymbol{\eta}(\boldsymbol{\theta})/\partial\theta_m$ are linearly independent for every $\boldsymbol{\theta} \in \Theta$. The vector $\mathbf{y} \in \mathbb{R}^N$ is observed, the mapping $\boldsymbol{\eta}$ and the set Θ are known, $\boldsymbol{\Sigma}$ is either known, or of the form $\boldsymbol{\Sigma} = c\mathbf{W}$ with $c > 0$ unknown and $\mathbf{W}$ known. Statistical inference on the unknown vector $\boldsymbol{\theta}$ should be performed.

A well known point estimator in model (1) is the maximum likelihood (= M. L.) estimator

$$\hat{\boldsymbol{\theta}} := \hat{\boldsymbol{\theta}}(\mathbf{y}) := \arg\min_{\boldsymbol{\theta}\in\Theta} \|\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta})\|^2_{\mathbf{W}}. \tag{2}$$

Here $\|\boldsymbol{a}\|^2_{\mathbf{W}} := \boldsymbol{a}^{\mathrm{T}}\mathbf{W}^{-1}\boldsymbol{a}$; $(\boldsymbol{a} \in R^N)$.

In the particular case when model (1) is linear, the statistic $\mathbf{y} \in \mathbb{R}^N \mapsto \hat{\boldsymbol{\theta}}(\mathbf{y})$ is not only a point estimator, it is also a sufficient statistic. If model (1) is nonlinear (more exactly, if the expectation surface of model (1)

$$\mathscr{E} := \{\boldsymbol{\eta}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

is not a "plane", the statistic $\hat{\boldsymbol{\theta}}(\mathbf{y})$ is no more sufficient. Consequently it contains less information about $\boldsymbol{\theta}$ than the sample vector $\mathbf{y}$. (For the distributional properties of $\hat{\boldsymbol{\theta}}(\mathbf{y})$ cf. e.g. [4, 5]).

However, it is possible to look for a statistic in model (1) which is a sufficient statistic, and which is somehow related to the M. L. estimator. In particular, we can require that this statistic coincides with $\hat{\boldsymbol{\theta}}(\mathbf{y})$ when model (1) is linear.

In Section 2 we propose such statistics. They have the following geometrical origin:

Consider the expectation surface $\mathscr{E}$. It is an m-dimensional surface in the N dimensional sample space $\mathbb{R}^N$. According to (2), the point $\boldsymbol{\eta}(\hat{\boldsymbol{\theta}}) \in \mathscr{E}$ is obtained by the $\mathbf{W}$-orthogonal projection of the point $\mathbf{y}$ onto $\mathscr{E}$. Consider now for any $\boldsymbol{\theta}^* \in \Theta$ the set

$$T_{\boldsymbol{\theta}^*} := \left\{\boldsymbol{\eta}(\boldsymbol{\theta}^*) + \frac{\partial \boldsymbol{\eta}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \mathbf{v} : \mathbf{v} \in \mathbb{R}^m\right\}.$$

Geometrically, $T_{\boldsymbol{\theta}^*}$ is the tangent plane to the surface $\mathscr{E}$ at the point $\boldsymbol{\eta}(\boldsymbol{\theta}^*) \in \mathscr{E}$. Statistically, $T_{\boldsymbol{\theta}^*}$ is the expectation surface of a linear model which approximates model (1):

(3) $$\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta}^*) = \frac{\partial \boldsymbol{\eta}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^{\mathrm{T}}} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} \sim \mathscr{N}(\mathbf{0}, \Sigma).$$

The M. L. estimate in this linearized model is

(4) $$\boldsymbol{\tau}(\mathbf{y}, \boldsymbol{\theta}^*) := \arg\min_{\boldsymbol{\theta}} \left\| \mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta}^*) - \frac{\partial \boldsymbol{\eta}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^{\mathrm{T}}} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \right\|_{\mathbf{W}}^2.$$

It is the result of the $\mathbf{W}$-orthogonal projection of the point $\mathbf{y}$ onto $T_{\boldsymbol{\theta}^*}$.

The statistic $\mathbf{y} \mapsto \boldsymbol{\tau}(\mathbf{y}, \boldsymbol{\theta}^*)$ is sufficient in model (3), however, it is not in model (1). Therefore, we proceed further by considering not one but many (eventually all) tangent planes to $\mathscr{E}$, and by projecting $\mathbf{W}$-orthogonally the sample point $\mathbf{y}$ onto all of them. (The reader which is familiar with differential geometry see that we are using the "tangent space" of $\mathscr{E}$). Consequently, instead of one random vector $\boldsymbol{\tau}(\mathbf{y}, \boldsymbol{\theta}^*)$ we consider the set of random vectors

(5) $$\{\boldsymbol{\tau}(\mathbf{y}, \boldsymbol{\theta}^*);\ \boldsymbol{\theta}^* \in D\}$$

for some $D \subset \Theta$. Evidently, this is a (vector-valued) random process defined on D. This process will be shown to have several pleasant structural properties.

a) It is a Gaussian random process having a covariance function which does not depend on $\boldsymbol{\theta}$.

b) Each component $\tau(\mathbf{y}, \boldsymbol{\theta}^*)$ of this process is related to a linear approximative model.

c) When D is adequately chosen, the mapping

$$\mathbf{y} \in \mathbb{R}^N \mapsto \{\tau(\mathbf{y}, \boldsymbol{\theta}^*);\ \boldsymbol{\theta}^* \in D\}$$

is a sufficient statistic in model (1).

In Section 3–5 we try to demonstrate that such a process is useful. We restrict our attention to the case of a finite $D = \{\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^k\}$, and instead of the process we consider a $k \,.\, m$ dimensional random vector $\tau(\mathbf{y}) := (\tau^{\mathrm{T}}(\mathbf{y}, \boldsymbol{\theta}^1), \ldots, \tau^{\mathrm{T}}(\mathbf{y}, \boldsymbol{\theta}^k))^{\mathrm{T}}$. If we linearize each component of $\tau(\mathbf{y})$ separately, we obtain a new, nonstandard linearization of model (1) which is more efficient than the standard linearization (3) (see Proposition 2). This allows to obtain an approximative expression for the posterior probability density of $\boldsymbol{\theta}$ (Proposition 3). Moreover, using quadratic functions of $\tau(\mathbf{y})$ we can discuss some confidence regions for $\boldsymbol{\theta}$, both for the case when $\boldsymbol{\Sigma}$ is known and when $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$ with an unknown σ.

2. SUFFICIENT STATISTICS

As is well known (cf. [1], Chapt. VIII.1.), the M.L. estimate of $\boldsymbol{\theta}$ in the linear model (3) can be expressed in the form

$$\tau(\mathbf{y}, \boldsymbol{\theta}^*) = \mathbf{M}^{-1}(\boldsymbol{\theta}^*)\, \mathbf{F}^{\mathrm{T}}(\boldsymbol{\theta}^*)\, \mathbf{W}^{-1}[\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta}^*)] + \boldsymbol{\theta}^* \tag{6}$$

where

$$\{\mathbf{F}(\boldsymbol{\theta})\}_{ij} := \frac{\partial \eta_i(\boldsymbol{\theta})}{\partial \theta_j}; \quad (i = 1, \ldots, N, \quad j = 1, \ldots, m),$$

$$\mathbf{M}(\boldsymbol{\theta}) := \mathbf{F}^{\mathrm{T}}(\boldsymbol{\theta})\, \mathbf{W}^{-1}\, \mathbf{F}(\boldsymbol{\theta}).$$

Consequently, (5) is a Gaussian random process with the mean

$$\boldsymbol{m}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^*) = \mathbf{M}^{-1}(\boldsymbol{\theta}^*)\, \mathbf{F}^{\mathrm{T}}(\boldsymbol{\theta}^*)\, \mathbf{W}^{-1}[\boldsymbol{\eta}(\boldsymbol{\theta}) - \boldsymbol{\eta}(\boldsymbol{\theta}^*)] + \boldsymbol{\theta}^*; \quad (\boldsymbol{\theta}^* \in D) \tag{7}$$

and with the covariance function $c\mathbf{K}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^0)$ where

$$\mathbf{K}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^0) = \mathbf{M}^{-1}(\boldsymbol{\theta}^*)\, \mathbf{F}^{\mathrm{T}}(\boldsymbol{\theta}^*)\, \mathbf{W}^{-1}\, \mathbf{F}(\boldsymbol{\theta}^0)\, \mathbf{M}^{-1}(\boldsymbol{\theta}^0).$$

We see that $\mathbf{K}(\boldsymbol{\theta}^*, \boldsymbol{\theta}^0)$ does not depend on $\boldsymbol{\theta}$ (= the true value of the parameters).

When the set D is finite, $D = \{\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^k\}$, it is better to consider the $(m \,.\, k)$-dimensional random vector

$$\boldsymbol{\tau} := \tau(\mathbf{y}) := \begin{pmatrix} \tau(\mathbf{y}, \boldsymbol{\theta}^1) \\ \vdots \\ \tau(\mathbf{y}, \boldsymbol{\theta}^k) \end{pmatrix}, \tag{8}$$

instead of the random process (5). Here each component $\tau(\mathbf{y}, \boldsymbol{\theta}^i)$ is defined according to (6).

The mean and the variance matrix of τ are equal to

$$\boldsymbol{m}_\theta := \mathsf{E}_\theta(\tau) = (\boldsymbol{m}_\theta^{\mathrm{T}}(\theta^1), \ldots, \boldsymbol{m}_\theta^{\mathrm{T}}(\theta^k))^{\mathrm{T}}$$

$$\mathsf{Var}\,(\tau) = c\,\mathbf{S}$$

where

(9)
$$\mathbf{S} := \begin{pmatrix} \mathbf{K}(\theta^1, \theta^1), \ldots, \mathbf{K}(\theta^1, \theta^k) \\ \mathbf{K}(\theta^k, \theta^1), \ldots, \mathbf{K}(\theta^k, \theta^k) \end{pmatrix}.$$

If $\mathbf{A}$ is any $r \times s$ matrix, we denote by $\mathscr{M}(\mathbf{A}) := \{\mathbf{A}\boldsymbol{u} : \boldsymbol{u} \in \mathbb{R}^s\}$ the linear subspace of $\mathbb{R}^r$ spanned by the columns of $\mathbf{A}$.

Proposition 1. If for every $\theta \in \Theta$

$$\mathscr{M}[\mathbf{F}(\theta)] \subset \mathscr{M}[(\mathbf{F}(\theta^1), \ldots, \mathbf{F}(\theta^k))] ,$$

then the statistic

$$\mathbf{y} \in \mathbb{R}^N \mapsto \tau(\mathbf{y}) \in \mathbb{R}^{mk}$$

is sufficient in model (1).

Proof. Let $\mathscr{L}$ be the linear manifold in $\mathbb{R}^N$ (the "plane") spanned by the set

$$\bigcup_{\theta \in \Theta} T_\theta .$$

Let us define

$$\mathbf{z}^\wedge := \mathbf{z}^\wedge(\mathbf{y}) := \arg\min_{\mathbf{z} \in \mathscr{L}} \|\mathbf{y} - \mathbf{z}\|_{\mathbf{W}}^2$$

The probability density of $\mathbf{y}$ is equal to

$$f(\mathbf{y} \mid \theta) = (2\pi)^{-N/2} \det^{-1/2}(\boldsymbol{\Sigma}) \exp\{-\|\mathbf{y} - \eta(\theta)\|_{\mathbf{W}}^2/(2c)\} \approx$$
$$\approx \exp\{-\|\mathbf{y} - \mathbf{z}^\wedge(\mathbf{y})\|_{\mathbf{W}}^2/(2c)\} \exp\{-\|\mathbf{z}^\wedge - \eta(\theta)\|_{\mathbf{W}}^2/(2c\} \; ; \quad (\theta \in \Theta) .$$

Hence, according to the factorisation theorem (cf. [1], Chapt. XV. 5.), the statistic $\mathbf{z}^\wedge(\mathbf{y})$ is sufficient in model (1).

Denote by

$$\mathbf{P}_i := \mathbf{F}(\theta^i)\, \mathbf{M}^{-1}(\theta^i)\, \mathbf{F}^{\mathrm{T}}(\theta^i)\, \mathbf{W}^{-1}$$

the $\mathbf{W}$-orthogonal projector onto $\mathscr{M}[\mathbf{F}(\theta^i)]$. The mapping $\mathbf{z} \mapsto (\mathbf{P}_1(\mathbf{z} - \eta(\theta^1)), \ldots \ldots, \mathbf{P}_k(\mathbf{z} - \eta(\theta^k))$ is one-to one on $\mathscr{L}$. Indeed, take $\mathbf{z}, \mathbf{z}^* \in \mathscr{L}$ such that

$$\mathbf{P}_i(\mathbf{z} - \eta(\theta^i)) = \mathbf{P}_i(\mathbf{z}^* - \eta(\theta^i)) \; ; \quad (i = 1, \ldots, k) .$$

Multiplying by $\mathbf{F}^{\mathrm{T}}(\theta^i)\, \mathbf{W}^{-1}$ from the left, we obtain

$$\mathbf{F}^{\mathrm{T}}(\theta^i)\, \mathbf{W}^{-1}(\mathbf{z} - \mathbf{z}^*) = 0 \; ; \quad (i = 1, \ldots, k) ,$$

i.e. $(\mathbf{z} - \mathbf{z}^*)$ is $\mathbf{W}$-orthogonal to $\mathscr{M}[(\mathbf{F}(\theta^1), \ldots, \mathbf{F}(\theta^k))]$, hence to $\mathscr{L}$. Consequently, $(\mathbf{z} - \mathbf{z}^*)^{\mathrm{T}}\, \mathbf{W}^{-1}(\mathbf{z} - \mathbf{z}^*) = 0$ hence, $\mathbf{z} - \mathbf{z}^*$.

It follows that $\mathbf{y} \in \mathbb{R}^N \mapsto (\mathbf{P}_1(\mathbf{z}^\wedge(\mathbf{y}) - \eta(\theta^1)), \ldots, \mathbf{P}_k(\mathbf{z}^\wedge(\mathbf{y}) - \eta(\theta^k)))$ is a sufficient statistic in model (1).

Since $\mathbf{z}^\wedge(\mathbf{y})$ is the $\mathbf{W}$-orthogonal projection of $\mathbf{y}$ onto $\mathscr{L}$ we have

$$\mathbf{P}_i(\mathbf{z}^\wedge(\mathbf{y}) - \eta(\theta^i)) = \mathbf{P}_i(\mathbf{y} - \eta(\theta^i)) \; ; \quad (i = 1, \ldots, k) .$$

Further, the equality

$$\mathbf{F}(\boldsymbol{\theta}^i)\,\tau(\mathbf{y}, \boldsymbol{\theta}^i) = \mathbf{P}_i(\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta}^i)) + \mathbf{F}(\boldsymbol{\theta}^i)\,\boldsymbol{\theta}^i$$

which follows from Eq. (6), specifies $\tau(\mathbf{y}, \boldsymbol{\theta}^i)$ uniquely, since $\mathbf{F}(\boldsymbol{\theta}^i)$ is of full rank. Consequently the mapping $\tau(\mathbf{y}) \mapsto (\mathbf{P}_1(\mathbf{z}^\wedge(\mathbf{y}) - \boldsymbol{\eta}(\boldsymbol{\theta}^1)), \ldots, \mathbf{P}_k(\mathbf{z}^\wedge(\mathbf{y}) - \boldsymbol{\eta}(\boldsymbol{\theta}^k))$ is one-to-one. It follows that $\tau(\mathbf{y})$ is a sufficient statistic in model (1). □

Corollary 1. If $D \subset \Theta$ is such that

$$\mathscr{M}[\mathbf{F}(\boldsymbol{\theta})] \subset \mathscr{M}[(\mathbf{F}(\boldsymbol{\theta}^1), \ldots, \mathbf{F}(\boldsymbol{\theta}^k))] \,; \quad (\boldsymbol{\theta} \in \Theta)$$

for some finite set $\{\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^k\} \subset D$, then

$$\mathbf{y} \in \mathbb{R}^N \mapsto \{\tau(\mathbf{y}, \boldsymbol{\theta}^*) \,; \quad \boldsymbol{\theta}^* \in D\}$$

is sufficient. Particularly

$$\mathbf{y} \in \mathbb{R}^N \mapsto \{\tau(\mathbf{y}, \boldsymbol{\theta}^*); \boldsymbol{\theta}^* \in \Theta\}$$

is always sufficient.

Corollary 2. Let $\pi(\boldsymbol{\theta})$ be a probability density on Θ (the prior density) such that

$$\mathscr{M}[\mathbf{F}(\boldsymbol{\theta})] \subset \mathscr{M}[(\mathbf{F}(\boldsymbol{\theta}^1), \ldots, \mathbf{F}(\boldsymbol{\theta}^k))] \,; \quad (\boldsymbol{\theta} \in \operatorname{supp}(\pi)) \,.$$

Then

$$\pi(\boldsymbol{\theta} \mid \tau(\mathbf{y})) = \pi(\boldsymbol{\theta} \mid \mathbf{y})$$

where $\pi(\boldsymbol{\theta} \mid \boldsymbol{u})$ denotes the posterior density of $\boldsymbol{\theta}$ given $\boldsymbol{u}$.

Proof. As in Proposition 1, we can prove that $\tau(\mathbf{y})$ is sufficient in the model

$$\mathbf{y} = \boldsymbol{\eta}(\boldsymbol{\theta}) + \varepsilon \,; \quad (\boldsymbol{\theta} \in \operatorname{supp}(\pi)) \,.$$

Hence $f(\mathbf{y} \mid \boldsymbol{\theta})$ can be factorized, i.e. we can write

$$f(\mathbf{y} \mid \boldsymbol{\theta}) = h(\mathbf{y})\, g(\tau(\mathbf{y}), \boldsymbol{\theta})$$

for some functions h and g. It follows that

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{f(\mathbf{y} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})}{\int_{\operatorname{supp}(\pi)} f(\mathbf{y} \mid \mathbf{t})\, \pi(\mathbf{t})\, \mathrm{d}\mathbf{t}} = \frac{g(\tau(\mathbf{y}), \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})}{\int_{\operatorname{supp}(\pi)} g(\tau(\mathbf{y}), \mathbf{t})\, \pi(\mathbf{t})\, \mathrm{d}\mathbf{t}} \,.$$

Hence $\mathbf{y} \in \mathbb{R}^N \mapsto \pi(\boldsymbol{\theta} \mid \mathbf{y})$ is a function of $\tau(\mathbf{y})$. According to the definition of conditional distributions (cf. [7], Chapt. V. 1.) it means that $\pi(\boldsymbol{\theta} \mid \mathbf{y}) = \pi(\boldsymbol{\theta} \mid \tau(\mathbf{y}))$. □

3. A NONSTANDARD LINEARIZATION

Let us consider the random vector $\tau(\mathbf{y})$ (the sufficient statistic) defined in Eq. (8). We have

(10) $$\tau(\mathbf{y}) \sim \mathscr{N}(\boldsymbol{m}_{\boldsymbol{\theta}}, c\, \mathbf{S}) \,; \quad (\boldsymbol{\theta} \in \Theta)$$

where $\boldsymbol{m}_{\boldsymbol{\theta}}$ and $\mathbf{S}$ are given by (9).

Instead of taking the linearization (3) we propose to linearize (10), i.e. to take approximatively

(11) $$\tau(\mathbf{y}) \sim \mathcal{N}(\mathbf{J}\boldsymbol{\theta}, c\,\mathbf{S})\,; \quad (\boldsymbol{\theta} \in \mathbb{R}^m)$$

where

$$\mathbf{J} := (\mathbf{I}, \ldots, \mathbf{I})^{\mathrm{T}}$$

and $\mathbf{I}$ is the $m \times m$ identity matrix. The linearization (11) is the linearization (3) applied separately to each component $\tau(\mathbf{y}, \boldsymbol{\theta}^i)$ of the vector $\tau(\mathbf{y})$.

To compare the standard linearization (3) with (11) take for $\boldsymbol{\theta}^*$ any point of the set $\{\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^k\}$, say $\boldsymbol{\theta}^* = \boldsymbol{\theta}^1$. Then consider the BLUE-s (= best linear unbiased estimates) of $\boldsymbol{\theta}$ in both models. The BLUE in model (3) is equal to $\tau(\mathbf{y}, \boldsymbol{\theta}^1)$, and is expressed in Eq. (6). Althoug the matrix $\mathbf{S}$ is singular (in general), and $\mathscr{M}(\mathbf{J}) \not\subset \mathscr{M}(\mathbf{S})$, the vector $\boldsymbol{\theta}$ can be estimated without bias in model (11), say by the estimate

$$\frac{1}{k}\,\mathbf{J}^{\mathrm{T}}\,\tau(\mathbf{y})\,.$$

Hence the BLUE exists also in model (11). Let us denote it by $\tilde{\boldsymbol{\theta}}(\mathbf{y})$. We refere to [3], Theorems 5.2.2 and 5.2.5 for explicit expressions for $\tilde{\boldsymbol{\theta}}(\mathbf{y})$ and $\operatorname{Var}\tilde{\boldsymbol{\theta}}(\mathbf{y})$. We have

$$\tilde{\boldsymbol{\theta}}(\mathbf{y}) = \mathbf{Q}\,\tau(\mathbf{y})\,, \quad \operatorname{Var}\tilde{\boldsymbol{\theta}}(\mathbf{y}) = c\,\mathbf{V}$$

where

(12) $$\mathbf{Q} := [\mathbf{J}^{\mathrm{T}}(\mathbf{S} + \mathbf{J}\mathbf{J}^{\mathrm{T}})^{-}\,\mathbf{J}]^{-1}\,\mathbf{J}^{\mathrm{T}}(\mathbf{S} + \mathbf{J}\mathbf{J}^{\mathrm{T}})^{-}$$
$$\mathbf{V} := [\mathbf{J}^{\mathrm{T}}(\mathbf{S} + \mathbf{J}\mathbf{J}^{\mathrm{T}})^{-}\,\mathbf{J}]^{-1} - \mathbf{I}$$

We note that $\mathbf{J}^{\mathrm{T}}(\mathbf{S} + \mathbf{J}\mathbf{J}^{\mathrm{T}})^{-}\,\mathbf{J}$ is nonsingular, since $\mathbf{J}$ is of full rank and $\mathscr{M}[\mathbf{J}] = \mathscr{M}[\mathbf{J}\mathbf{J}^{\mathrm{T}}] \subset \mathscr{M}[\mathbf{S} + \mathbf{J}\mathbf{J}^{\mathrm{T}}]$. In the particular case that $\mathbf{S}$ is regular, we have simpler formulae

(13) $$\mathbf{Q} = (\mathbf{J}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{J})^{-1}\,\mathbf{J}^{\mathrm{T}}\mathbf{S}^{-1}$$
$$\mathbf{V} = (\mathbf{J}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{J})^{-1}$$

Hence in the linearized model (11) we have

(14) $$\tilde{\boldsymbol{\theta}}(\mathbf{y}) \sim \mathcal{N}(\boldsymbol{\theta}, c\,\mathbf{V})\,; \quad (\boldsymbol{\theta} \in \mathbb{R}^m)$$

but in the linearized model (3) we have

(15) $$\tau(\mathbf{y}, \boldsymbol{\theta}^1) \sim \mathcal{N}(\boldsymbol{\theta}, c\,\mathbf{M}^{-1}(\boldsymbol{\theta}^1))\,; \quad (\boldsymbol{\theta} \in \mathbb{R}^m)\,.$$

To compare what linearization is better, we shall compare the exact distributions of $\tilde{\boldsymbol{\theta}}(\mathbf{y})$ and $\tau(\mathbf{y}, \boldsymbol{\theta}^1)$.

Proposition 2. The random vectors $\tilde{\boldsymbol{\theta}}(\mathbf{y})$ and $\tau(\mathbf{y}, \boldsymbol{\theta}^1)$ are exactly distributed according to

(16) $$\tilde{\boldsymbol{\theta}}(\mathbf{y}) \sim \mathcal{N}(\mathbf{Q}\mathbf{m}_{\boldsymbol{\theta}}, c\,\mathbf{V})\,; \quad (\boldsymbol{\theta} \in \Theta)$$

(17) $$\tau(\mathbf{y}, \boldsymbol{\theta}^1) \sim \mathcal{N}(\mathbf{m}_{\boldsymbol{\theta}}(\boldsymbol{\theta}^1)\,, c\,\mathbf{M}^{-1}(\boldsymbol{\theta}^1))\,; \quad (\boldsymbol{\theta} \in \Theta)\,.$$

The vectors expressing the bias

$$\mathbf{Q}\mathbf{m}_{\boldsymbol{\theta}} - \boldsymbol{\theta}$$

and the bias

$$\boldsymbol{m}_{\theta}(\boldsymbol{\theta}^1) - \boldsymbol{\theta}$$

are of the sample order of magnitude. The estimator $\tilde{\boldsymbol{\theta}}(\mathbf{y})$ is more efficient since the matrix $\mathrm{Var}\left[\boldsymbol{\tau}(\mathbf{y}, \boldsymbol{\theta}^1)\right] - \mathrm{Var}\left[\tilde{\boldsymbol{\theta}}(\mathbf{y})\right]$ is positive semidefinite.

Proof. Both variables $\tilde{\boldsymbol{\theta}}(\mathbf{y})$ and $\boldsymbol{\tau}(\mathbf{y}, \boldsymbol{\theta}^1)$ are linear in $\mathbf{y}$, hence they are normally distributed. The mean and the variance of $\boldsymbol{\tau}(\mathbf{y}, \boldsymbol{\theta}^1)$ is given in Eq. (7). The mean and the variance of $\tilde{\boldsymbol{\theta}}(\mathbf{y})$ follow from Eq. (12) and from the mean and the variance of $\boldsymbol{\tau}(\mathbf{y})$ in Eq. (9).

The bias of $\tilde{\boldsymbol{\theta}}(\mathbf{y})$ is

$$\mathbf{Q}\boldsymbol{m}_{\theta} - \boldsymbol{\theta} = \mathbf{Q}\begin{pmatrix} \boldsymbol{m}_{\theta}(\boldsymbol{\theta}^1) \\ \vdots \\ \boldsymbol{m}_{\theta}(\boldsymbol{\theta}^k) \end{pmatrix} - \boldsymbol{\theta}\,.$$

The bias of $\boldsymbol{\tau}(\mathbf{y}, \boldsymbol{\theta}^1)$ can be written in the form

$$\boldsymbol{m}_{\theta}(\boldsymbol{\theta}^1) - \boldsymbol{\theta} = \mathbf{QJ}\, \boldsymbol{m}_{\theta}(\boldsymbol{\theta}^1) - \boldsymbol{\theta} = \mathbf{Q}\begin{pmatrix} \boldsymbol{m}_{\theta}(\boldsymbol{\theta}^1) \\ \vdots \\ \boldsymbol{m}_{\theta}(\boldsymbol{\theta}^1) \end{pmatrix} - \boldsymbol{\theta}$$

since $\mathbf{QJ} = \mathbf{I}$, according to (12) and (13). Thus if $\boldsymbol{m}_{\theta}(\boldsymbol{\theta}^i) - \boldsymbol{\theta}$ is of the same order for every $i = 1, \ldots, k$, then $\boldsymbol{\tau}(\mathbf{y}, \boldsymbol{\theta}^1)$ and $\tilde{\boldsymbol{\theta}}(\mathbf{y})$ have the bias of the same order as well.

The random variable $\boldsymbol{\tau}(\mathbf{y}, \boldsymbol{\theta}^1)$ can be written in the form

$$\boldsymbol{\tau}(\mathbf{y}, \boldsymbol{\theta}^1) = (\mathbf{I}, \mathbf{0}, \ldots, \mathbf{0})\, \boldsymbol{\tau}(\mathbf{y})$$

hence it is a linear unbiassed estimator of $\boldsymbol{\theta}$ in model (11). Since $\tilde{\boldsymbol{\theta}}(\mathbf{y})$ is the BLUE in the same model, it follows that $\mathrm{Var}\left[\boldsymbol{\tau}(\mathbf{y}, \boldsymbol{\theta}^1)\right] - \mathrm{Var}\left[\tilde{\boldsymbol{\theta}}(\mathbf{y})\right]$ is positive semidefinite. □

Note 1. According to Eq. (7) we can write the bias in the form

$$\boldsymbol{m}_{\theta}(\boldsymbol{\theta}^1) - \boldsymbol{\theta} := \mathbf{r}(\boldsymbol{\theta}, \boldsymbol{\theta}^1)\,, \quad \mathbf{J}\boldsymbol{m}_{\theta} - \boldsymbol{\theta} = \mathbf{J}\begin{pmatrix} \mathbf{r}(\boldsymbol{\theta}, \boldsymbol{\theta}^1) \\ \vdots \\ \mathbf{r}(\boldsymbol{\theta}, \boldsymbol{\theta}^k) \end{pmatrix}$$

where from the Taylor formula for $\boldsymbol{\eta}(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^i$ we obtain

$$(18) \qquad \mathbf{r}(\boldsymbol{\theta}, \boldsymbol{\theta}^i) := \mathbf{M}^{-1}(\boldsymbol{\theta}^i)\, \mathbf{F}^{\mathrm{T}}(\boldsymbol{\theta}^i)\, \mathbf{W}^{-1}[\boldsymbol{\eta}(\boldsymbol{\theta}) - \boldsymbol{\eta}(\boldsymbol{\theta}^i)] + \boldsymbol{\theta}^i - \boldsymbol{\theta} =$$

$$= \tfrac{1}{2}\mathbf{M}^{-1}(\boldsymbol{\theta}^i)\, \mathbf{F}^{\mathrm{T}}(\boldsymbol{\theta}^i)\, \mathbf{W}^{-1}\left[(\boldsymbol{\theta} - \boldsymbol{\theta}^i)^{\mathrm{T}}\left[\frac{\partial^2 \boldsymbol{\eta}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^{\mathrm{T}}}\bigg|_{\lambda\boldsymbol{\theta} + (1-\lambda)\boldsymbol{\theta}^i}\right](\boldsymbol{\theta} - \boldsymbol{\theta}^i)\right]$$

for some number $\lambda \in (0, 1)$ depending on $\boldsymbol{\theta}$ and on $\boldsymbol{\theta}^i$.

The expression for $\mathbf{r}(\boldsymbol{\theta}, \boldsymbol{\theta}^i)$ is small either if $[\boldsymbol{\theta} - \boldsymbol{\theta}^i]^{\mathrm{T}}\, \mathbf{M}(\boldsymbol{\theta}^i)\, [\boldsymbol{\theta} - \boldsymbol{\theta}^i]$ is small or if model (1) is not too much curved, since

$$\sup\left\{\left\|\mathbf{v}^{\mathrm{T}}\, \frac{\partial^2 \boldsymbol{\eta}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^{\mathrm{T}}}\, \mathbf{v}\right\|_{\mathbf{W}} \Big/ \mathbf{v}^{\mathrm{T}}\, \mathbf{M}(\boldsymbol{\theta}^1)\, \mathbf{v}\,;\quad 0 \neq \mathbf{v} \in \mathbb{R}^m\right\}$$

is related to the curvatures of Bates and Watts [2] in model (1). We used here the notation

$$\mathbf{v}^{\mathrm{T}} \frac{\partial^2 \boldsymbol{\eta}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^{\mathrm{T}}} \mathbf{v} := \sum_{ij} v_i \frac{\partial^2 \boldsymbol{\eta}(\boldsymbol{\theta})}{\partial \theta_i\, \partial \theta_j} v_j .$$

It is important to note here that $\mathsf{E}_{\boldsymbol{\theta}}[\tilde{\theta}(\mathbf{y})] = \mathbf{Q}\boldsymbol{m}_{\boldsymbol{\theta}}$ is a "mixture" of the means of $\tau(\mathbf{y}, \boldsymbol{\theta}^1), \ldots, \tau(\mathbf{y}, \tilde{\boldsymbol{\theta}}^k)$. In some cases the "mixture" is such that the bias of $\tilde{\theta}(\mathbf{y})$ is much smaller than the bias of every $\tau(\mathbf{y}, \boldsymbol{\theta}^i)$. This depends on the choice off $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^k$.

Note 2. When $\tau(\mathbf{y})$ is a sufficient statistic (Proposition 1) we arrive to $\tilde{\theta}(\mathbf{y})$ according to the scheme

$$(1) \mapsto \text{sufficient statistic } \boldsymbol{\tau} \mapsto (10) \mapsto \text{linearization of } \boldsymbol{\tau} \mapsto (11) \quad \mapsto$$
$$\mapsto \text{sufficient statistic } \tilde{\boldsymbol{\theta}} \mapsto (14)$$

Example 1. We shall consider the simple nonlinear model $\mathbf{y} = \boldsymbol{\eta}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}$ with $N = 2$, $m = 1$, $\Theta = (0, \pi)$, $\boldsymbol{\eta}(\boldsymbol{\theta}) = (\cos \theta, \sin \theta)^{\mathrm{T}}$, $\boldsymbol{\Sigma} = \mathbf{W} = \mathbf{I}$. (The expectation surface is a halfcircle). In this case we have $\mathbf{F}^{\mathrm{T}}(\theta) = (-\sin \theta, \cos \theta)$, $\mathbf{M}(\theta) = 1$; $(\theta \in \Theta)$. To construct $\tau(\mathbf{y})$ take two points $\theta^1 = \theta^* - \delta$, $\theta^2 = \theta^* + \delta$ for some fixed $\delta > 0$, $\theta^* \in \Theta$. By simple computations we obtain

$$\tau(\mathbf{y}) = \begin{pmatrix} -y_1 \sin(\theta^* - \delta) + y_2 \cos(\theta^* - \delta) + \theta^* - \delta \\ -y_1 \sin(\theta^* + \delta) + y_2 \cos(\theta^* + \delta) + \theta^* + \delta \end{pmatrix}$$

$$\mathbf{S} = \begin{pmatrix} 1, & \cos(2\delta) \\ \cos(2\delta), & 1 \end{pmatrix}, \quad \mathbf{S}^{-1} = \begin{pmatrix} 1, & -\cos(2\delta) \\ -\cos(2\delta), & 1 \end{pmatrix} \Big/ \sin^2(2\delta)$$

$$\mathbf{V} = (\mathbf{J}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{J})^{-1} = (1 + \cos 2\delta)/2 = \cos^2 \delta < 1$$

$$\tilde{\theta}(\mathbf{y}) = (\mathbf{J}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{J})^{-1}\, \mathbf{J}^{\mathrm{T}}\mathbf{S}^{-1}\, \tau(\mathbf{y}) = \cos \delta[-y_1 \sin \theta^* + y_2 \cos \theta^*] + \theta^*$$

When $\delta \mapsto 0$ we obtain $\tau(\mathbf{y}, \theta^*)$:

$$\tau(\mathbf{y}, \theta^*) = [-y_1 \sin \theta^* + y_2 \cos \theta^*] + \theta^* .$$

Further

$$\mathsf{E}_\theta[\tilde{\theta}(\mathbf{y})] - \theta = \cos \delta \sin(\theta - \theta^*) + (\theta^* - \theta),$$
$$\mathsf{E}_\theta[\tau(\mathbf{y}, \theta^*)] - \theta = \sin(\theta - \theta^*) + (\theta^* - \theta).$$

Hence for δ not very large, the bias of $\tilde{\theta}(\mathbf{y})$ and of $\tau(\mathbf{y}, \theta^*)$ is approximatively the same. The mean square error of $\tilde{\theta}(\mathbf{y})$ is equal to

$$\mathsf{E}_\theta[\tilde{\theta}(\mathbf{y}) - \theta]^2 = \cos^2 \delta + [\cos \delta \sin(\theta - \theta^*) + (\theta^* - \theta)]^2 := \psi(\delta).$$

We have

$$\frac{\mathrm{d}\psi}{\mathrm{d}\delta} = -2 \cos \delta \sin \delta [1 + \sin^2(\theta - \theta^*)] - 2 \sin \delta [\sin(\theta - \theta^*)](\theta^* - \theta).$$

Hence $d\psi/d\delta|_{\delta=0} = 0$. Further

$$\left.\frac{d^2\psi}{d\delta^2}\right|_{\delta=0} = -2[-\sin^2\delta + \cos^2\delta][1 + \sin^2(\theta - \theta^*)]|_{\delta=0} -$$

$$- 2\cos\delta[\sin(\theta - \theta^*)](\theta^* - \theta)|_{\delta=0} =$$

$$= -2[1 + \sin^2(\theta - \theta^*)] + 2(\theta - \theta^*)\sin(\theta - \theta^*).$$

If $\theta > \theta^*$ then

$$\left.\frac{d^2\psi}{d\delta^2}\right|_{\delta=0} \leqq -2 + 2[\theta - \theta^* - \sin(\theta - \theta^*)].$$

If $\theta < \theta^*$ then

$$\left.\frac{d^2\psi}{d\delta^2}\right|_{\delta=0} \leqq -2 + 2[\theta^* - \theta - \sin(\theta^* - \theta))].$$

Hence, if θ^* is so near to θ that

$$|\theta - \theta^*| - \sin(|\theta - \theta^*|) < 1$$

then $\mathsf{E}_\theta[\tilde{\theta}(\mathbf{y}) - \theta]^2$ attains its maximum at $\delta = 0$. Consequently

$$\mathsf{E}_\theta[\tilde{\theta}(\mathbf{y}) - \theta]^2 < \mathsf{E}_\theta[\tau(\mathbf{y}, \theta^*) - \theta]^2.$$

4. THE POSTERIOR PROBABILITY DENSITY OF $\boldsymbol{\theta}$

Consider a normal prior density $\pi(\boldsymbol{\theta})$ in model (1),

$$\pi(\boldsymbol{\theta}) = (2\pi)^{-m/2} \det^{-1/2}(\mathbf{H}) \exp\{-\tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^0)^T \mathbf{H}^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^0)\},$$

where $\mathbf{H}$ is a given matrix and $\boldsymbol{\theta}^0 \in \Theta$ is a given vector. Denote by $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ the corresponding posterior density. If $\boldsymbol{\tau}(\mathbf{y})$ is a sufficient static (Corollary 2 to Proposition 1) then

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) = \pi(\boldsymbol{\theta} \mid \boldsymbol{\tau}(\mathbf{y})).$$

This is not a normal density. However, using the linearization described in Section 3, we can write approximatively

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) \doteq \pi_{\text{lin}}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}(\mathbf{y}))$$

where $\tilde{\boldsymbol{\theta}}(\mathbf{y})$ is supposed to be distributed according to Eq. (14).

Proposition 3. $\pi_{\text{lin}}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}(\mathbf{y}))$ is a normal probability density with the mean equal to

$$\boldsymbol{\theta}^0 + \mathbf{H}(c\,\mathbf{V} + \mathbf{H})^{-1}(\tilde{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}^0) \tag{19}$$

and with the variance matrix equal to

$$\mathbf{H} - \mathbf{H}(c\,\mathbf{V} + \mathbf{H})^{-1}\mathbf{H} \tag{20}$$

where $\tilde{\boldsymbol{\theta}}(\mathbf{y})$ and $\mathbf{V} = \mathsf{Var}\,\tilde{\boldsymbol{\theta}}(\mathbf{y})$ are defined by Eqs. (12) resp. (13).

Proof. Denote by $h_{\text{lin}}(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\theta})$ the probability density of $\tilde{\boldsymbol{\theta}}$ corresponding to Eq. (14).

Consider the vector

$$\begin{pmatrix} \tilde{\boldsymbol{\theta}} \\ \boldsymbol{\theta} \end{pmatrix}$$

as a random vector with the joint density $\pi(\boldsymbol{\theta})\, h_{\text{lin}}(\tilde{\boldsymbol{\theta}} \mid \boldsymbol{\theta})$ and denote by $\mathsf{E}(\cdot)$ the operator of taking the mean with respect to this density. By simple computations we obtain

$$\begin{aligned}
&\mathsf{E}(\boldsymbol{\theta}) = \boldsymbol{\theta}^0 \\
&\mathsf{E}(\tilde{\boldsymbol{\theta}}) = \mathsf{E}[\mathsf{E}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}})] = \boldsymbol{\theta}^0 \\
&\mathsf{E}[(\boldsymbol{\theta} - \boldsymbol{\theta}^0)(\theta - \boldsymbol{\theta}^0)^{\mathrm{T}}] = \mathbf{H} \\
&\mathsf{E}[(\boldsymbol{\theta} - \boldsymbol{\theta}^0)(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)^{\mathrm{T}} = \mathsf{E}[(\boldsymbol{\theta} - \boldsymbol{\theta}^0)\, \mathsf{E}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)^{\mathrm{T}}] = \mathbf{H} \\
&\mathsf{E}[(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)^{\mathrm{T}}] = \mathsf{E}[\mathsf{E}_{\boldsymbol{\theta}}[(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)^{\mathrm{T}}]] = c\,\mathbf{V} + \mathbf{H}\,.
\end{aligned}$$

Hence

$$\begin{pmatrix} \tilde{\boldsymbol{\theta}} \\ \boldsymbol{\theta} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\theta}^0 \\ \boldsymbol{\theta}^0 \end{pmatrix}, \begin{pmatrix} c\,\mathbf{V} + \mathbf{H}, & \mathbf{H} \\ \mathbf{H}, & \mathbf{H} \end{pmatrix}\right).$$

According to [6], Chapt. 8. a 2, (V), the conditional density of $\boldsymbol{\theta}$ given $\tilde{\boldsymbol{\theta}}$ is normal with the mean

$$\boldsymbol{\theta}^0 + \mathbf{H}(c\,\mathbf{V} + \mathbf{H})^{-1}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)$$

and with the variance

$$\mathbf{H} - \mathbf{H}(c\,\mathbf{V} + \mathbf{H})^{-1}\,\mathbf{H}\,. \qquad \square$$

Note. The statistic $\tilde{\boldsymbol{\theta}}(\mathbf{y})$ is sufficient in the linearized model (11). Therefore we can write (compare with Corollary 2)

$$\pi_{\text{lin}}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}(\mathbf{y})) = \pi_{\text{lin}}(\boldsymbol{\theta} \mid \boldsymbol{\tau}(\mathbf{y}))\,.$$

On the other hand, the exact posterior density is

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) = \pi(\boldsymbol{\theta} \mid \boldsymbol{\tau}(\mathbf{y}))\,.$$

Hence we can compare the approximative and the exact posterior density from

$$\frac{\pi_{\text{lin}}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}(\mathbf{y}))}{\pi(\boldsymbol{\theta} \mid \mathbf{y})} = \frac{\pi_{\text{lin}}(\boldsymbol{\theta} \mid \boldsymbol{\tau}(\mathbf{y}))}{\pi(\boldsymbol{\theta} \mid \boldsymbol{\tau}(\mathbf{y}))}\,.$$

In $\pi_{\text{lin}}(\boldsymbol{\theta} \mid \boldsymbol{\tau}(\mathbf{y}))$ we take $\mathsf{E}_{\boldsymbol{\theta}}[\boldsymbol{\tau}(\mathbf{y})] = \mathbf{J}\boldsymbol{\theta}$, in $\pi(\boldsymbol{\theta} \mid \boldsymbol{\tau}(\mathbf{y}))$ we take $\mathsf{E}_{\boldsymbol{\theta}}[\boldsymbol{\tau}(\mathbf{y})] = \boldsymbol{m}_{\boldsymbol{\theta}}$, otherwise the Bayes formulae for computing $\pi_{\text{lin}}(\boldsymbol{\theta} \mid \boldsymbol{\tau}(\mathbf{y}))$ and $\pi(\boldsymbol{\theta} \mid \boldsymbol{\tau}(\mathbf{y}))$ are the same.

5. A NOTE ON CONFIDENCE REGIONS FOR $\boldsymbol{\theta}$

We consider confidence regions for $\boldsymbol{\theta}$ which are based on $\tilde{\boldsymbol{\theta}}(\mathbf{y})$. We note that they are of restricted importance, since they are influence by the choice of $\boldsymbol{\theta}^*$ in (3), resp. by the choice of the points $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^k$ in (8) which in fact represents a prior knowledge about $\boldsymbol{\theta}$.

From (16) we obtain that the set

$$\{\boldsymbol{\theta}: \|\tilde{\theta}(\mathbf{y}) - \mathbf{Qm}_{\theta}\|_{\mathbf{V}}^2 < c\chi_m^2(\alpha)\} \tag{21}$$

is a confidence region for $\boldsymbol{\theta}$ in the case that $c\mathbf{W}$ is known. α is the exact confidence level, and $\chi_m^2(\alpha)$ is the α-quantile of the χ^2 distribution with m degrees of freedom.

Example 2. Take the set-up from Example 1. We have

$$\mathbf{Qm}_{\theta} = \mathrm{E}_{\theta}[\tilde{\theta}(\mathbf{y})] = \cos\delta[-\cos\theta\sin\theta^* + \sin\theta\cos\theta^*] + \theta^* .$$

Hence

$$\|\tilde{\theta}(\mathbf{y}) - \mathbf{Qm}_{\theta}\|_{\mathbf{V}}^2 = [-(y_1 - \cos\theta)\sin\theta^* + (y_2 - \sin\theta)\cos\theta^*]^2 .$$

We see that the confidence region (21) does not depend on δ, hence the standard and the nonstandard linearizations are equivalent as regard to the confidence regions. This is by no way in contradiction to Proposition 2; the random variable $\tilde{\theta}(\mathbf{y})$ has a small variance, however, this has no importance for confidence reasoning. On the other hand, the obtained confidence region depends very much on θ^*.

To understand the situation geometrically, let us write $\tilde{\boldsymbol{\theta}}(\mathbf{y})$ in the form

$$\tilde{\boldsymbol{\theta}}(\mathbf{y}) = \mathbf{Ly} + \boldsymbol{l}$$

for some matrix $\mathbf{L}$ and some vector $\boldsymbol{l}$ (This is possible, since $\tilde{\theta}(\mathbf{y})$ is linear in $\mathbf{y}$). Further we have

$$c\,\mathbf{V} = \mathrm{Var}\,\tilde{\boldsymbol{\theta}}(\mathbf{y}) = c\,\mathbf{LW}^{-1}\mathbf{L}^{\mathrm{T}}$$

hence

$$\mathbf{P} := \mathbf{L}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{LW}^{-1}$$

is a $\mathbf{W}$-orthogonal projector. We can verify that

$$\|\mathbf{P}[\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta})]\|_{\mathbf{W}}^2 = \|\tilde{\boldsymbol{\theta}}(\mathbf{y}) - \mathbf{Qm}_{\theta}\|_{\mathbf{V}}^2 . \tag{22}$$

Hence the confidence region (21) has the form

$$\{\boldsymbol{\theta}: \|\mathbf{P}[\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta})]\|_{\mathbf{W}}^2 < c\,\chi_m^2(\alpha)\} .$$

This confidence region, although exact, gives poor results (it is too large) if the value $\|\mathbf{P}[\boldsymbol{\eta}(\boldsymbol{\theta}_{\mathrm{true}}) - \boldsymbol{\eta}(\boldsymbol{\theta}^*)]\|_{\mathbf{W}}^2$ is large. (We note, that this is zero if model (1) is linear.)

Another consequence of (22) is that $\|\tilde{\boldsymbol{\theta}}(\mathbf{y}) - \mathbf{Qm}_{\theta}\|_{\mathbf{V}}^2$ and $\|(\mathbf{I} - \mathbf{P})[\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta})]\|_{\mathbf{W}}^2$ are independent random variables. Hence another confidence region (of the exact confidence level α) is of the form

$$\left\{\boldsymbol{\theta}: \frac{(N-m)\,\|\tilde{\boldsymbol{\theta}}(\mathbf{y}) - \mathbf{Qm}_{\theta}\|_{\mathbf{V}}^2}{m\|(\mathbf{I} - \mathbf{P})[\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta})]\|_{\mathbf{W}}^2} < F_{m,N-m}(\alpha)\right\}$$

where $\mathrm{F}_{m,N-m}(\alpha)$ is the α-quantile of the F-distribution with m and $N - m$ degrees of freedom. The advantage of this region comparing with (21) is that it can be used in the case when c is unknown.

6. CONSEQUENCES FOR NONLINEAR EXPERIMENTAL DESIGN

The covariance matrix of $\tilde{\theta}(\mathbf{y})$ (Eq. (16)), and the approximative aposteriori covariance matrix (Eq. (20)) do not depend on the observed vector $\mathbf{y}$, and are smaller than the corresponding variances in the standard linearization. Therefore they are adequate to construct optimality criteria for optimum experimental design in nonlinear models.

(Received March 28, 1989.)

REFERENCES

[1] J. Anděl: Matematická statistika (Mathematical Statistics). SNTL/ALFA, Praha 1978.
[2] D. M. Battes and D. G. Watts: Relative curvature measures of nonlinearity. J. Roy. Statist. Soc. Ser. B *42* (1980), 1—25.
[3] L. Kubáček: Foundations of Estimation Theory. Elsevier, Amsterdam 1988. (Slovak edition: Základy teórie odhadu. Veda, Bratislava 1982.)
[4] A. Pázman: Probability distribution of the multivariate nonlinear least squares estimates. Kybernetika *20* (1984), 209—230.
[5] A. Pázman: On information matrices in nonlinear experimental design. J. Statist. Plann. Inference *21* (1989), 253—263.
[6] R. C. Rao: Linear Statistical Inference and Its Applications. J. Wiley, New York 1973. (Czech translation: Lineární metódy statistické indukce a jejich aplikace. Academia, Praha 1978.)
[7] A. Rényi: Teorie pravděpodobnosti. Academia, Praha 1972. (German original: Wahrscheinlichkeitsrechnung mit einem Anhang über Informationstheorie. VEB Deutscher Verlag d. Wissenschaften, Berlin 1962.)

Andrej Pázman, Matematický ústav SAV (Mathematical Institute — Slovak Academy of Sciences), Obrancov mieru 49, 814 73 Bratislava. Czechoslovakia.