INTRINSIC LINEARIZATION OF NONLINEAR REGRESSION BY PRINCIPAL COMPONENTS METHOD

K. HORNIŠOVÁ

Abstract. Most commonly nonlinear regression models have an important parameter-effect nonlinearity but only a small intrinsic nonlinearity. Hence it is of interest to approximate them linearly. This can be done either by retaining the original parametrization θ, or by choosing a new parametrization $\beta = \beta(\theta)$. Under a prior weight density $\pi(\theta)$ we propose criterion of optimality of intrinsically linear approximation. The optimal solution is obtained by principal components method. The distance of the expectation surface of the new model from the expectation surface of the original one can be considered as a measure of intrinsic nonlinearity of the original model, which is simpler to compute than the well-known measure of Bates and Watts (1980). In the examples consequences for inference on parameters are examined.

1. Introduction

We consider a (not necessary regular) nonlinear regression model

$$\begin{aligned} &y = \eta(\theta) + \varepsilon, \qquad \theta \in \Theta \subseteq \mathbb{R}^m, \\ &\varepsilon \sim N(0, \sigma^2 W), \end{aligned} \tag{1}$$

where $\eta(.) : \Theta \to \mathbb{R}^N$ is measurable mapping, $y \in \mathbb{R}^N$ is vector of measurements, $\varepsilon \in \mathbb{R}^N$ vector of random errors, W is known (positive definite) matrix (usually $W = I$), σ^2 is unknown.

Received January 7, 2004.

2000 *Mathematics Subject Classification.* Primary 62J02; Secondary 62F25.

Key words and phrases. Nonlinear regression, linearization, prior, principal components analysis.

Research supported by the Slovak Vega grant No. 2/4026/04.

Since the statistical inference in linear models is much more simpler than in nonlinear models, possibility of using some linear model instead of the model (1) and a proper choice of it is often studied. It is natural to base the linearization on some prior information on unknown parameter θ, if it is available, and then compare efficiency of such simplified methods with corresponding exact methods.

Situation, when it is known that the true value $\bar{\theta}$ of parameter θ lies in a neighbourhood of a given prior point $\theta^0 \in \Theta$, was treated most often. If the regression function of the original model is twice continuously differentiable, the model (1) is then linearized by the linear part of its Taylor expansion at θ^0, i.e. by the model

$$\begin{aligned} y &= \eta(\theta^0) + \frac{\partial \eta(\theta^0)}{\partial \theta^\top}(\theta - \theta^0) + \varepsilon = A\theta + a + \varepsilon, \\ \varepsilon &\sim N(0, \sigma^2 W). \end{aligned} \tag{2}$$

Similarly can be linearized arbitrary parametric function $g(\theta)$ of interest. Properties and conditions on admissibility of Taylor linearization in a prior point were studied e.g. in [4].

Sometimes, the prior distribution π on Θ is known. For this case, different ways of linearization of (1) were proposed. For one of the methods – nonstandard linearization, see [6].

Another way how to linearize the model (1) utilizing the knowledge of prior π is linearization by smoothing, proposed in [7]. The approximative linear model

$$\begin{aligned} Y &= A\theta + a + \varepsilon, \\ \varepsilon &\sim N(0, \sigma^2 W), \end{aligned} \tag{3}$$

is chosen according to the criterion

$$K_1 := \min_{\substack{A\in\mathbb{R}^{N\times m}\\ a\in\mathbb{R}^N}} \mathrm{E}_\pi[\|\eta(\theta)-(A\theta+a)\|_W^2]$$

$$(4) \qquad = \min_{\substack{A\in\mathbb{R}^{N\times m}\\ a\in\mathbb{R}^N}} \int_\Theta \|\eta(\theta)-(A\theta+a)\|_W^2 \pi(\theta)d\theta.$$

The solutions of the minimization problem (4), which corresponds to minimization of prior expectation of I-divergence between the nonlinear and linear model, have the form

$$(5) \qquad \begin{aligned} A &= \mathrm{Cov}_\pi(\eta,\theta)(\mathrm{Var}_\pi\theta)^-, \\ a &= \mathrm{E}_\pi\eta - A\,\mathrm{E}_\pi\theta, \\ K_1 &= \mathrm{tr}\ W^{-1/2}\{\mathrm{Var}_\pi\eta - \mathrm{Cov}_\pi(\eta,\theta)(\mathrm{Var}_\pi\theta)^-\,\mathrm{Cov}_\pi(\theta,\eta)\}W^{-1/2}, \end{aligned}$$

where the last expression is invariant with regard to the choice of pseudoinversion $(\mathrm{Var}_\pi\theta)^-$, if the indicated means and covariances with regard to prior distribution $\pi(.)$ on Θ exist and are finite. Parametric functions of interest can be linearized accordingly.

The advantage of the method is that it can be used also when the response function $\eta(\theta)$ has no derivatives. In the other case

$$A_{\pi_n} \longrightarrow \frac{\partial\eta(\theta^0)}{\partial\theta^\top}$$

if

$$\pi_n \rightarrow \pi_0,$$

where π_n are nondegenerate prior distributions and π_0 is distribution concentrated at θ^0.

In [3], under the knowledge of joint prior π for (θ,σ), the linearization of the model is circumvented and parametric function $g(\theta,\sigma)\in\mathbb{R}^s$ of interest is directly estimated by explicit estimator $\hat{g}(y)$, which is defined as linear combination $\mathcal{A}\Phi(y)$, $\mathcal{A} \in \mathbb{R}^{s\times u}$, of given functions $\Phi(.) : \mathbb{R}^n \to \mathbb{R}^u$ of observations, such that the coefficient $\mathcal{A}$ is optimal with regard to criterion (average mean square error (AMSE))

$$\min_{\mathcal{A}\in\mathbb{R}^{s\times u}} \mathrm{E}_{\pi} \mathrm{E}_{f(.|\theta,\sigma)} \|g(\theta) - \mathcal{A}\Phi(y)\|^2, \tag{6}$$

where $f(.|\theta,\sigma)$ is conditional density of y.

For example, if $\Phi(y) = \begin{pmatrix} y \\ 1 \end{pmatrix}$, then linear explicit estimator is $\mathcal{A}\Phi(y) = A^\top y + a$, where

$$\begin{aligned} A &= [\mathrm{Var}_{\pi}\eta + \mathrm{E}_{\pi}(\sigma^2)W]^{-1} \mathrm{Cov}_{\pi}(\eta, g), \\ &\text{and} \\ a &= \mathrm{E}_{\pi} g - A^\top \mathrm{E}_{\pi}\eta, \end{aligned} \tag{7}$$

with

$$\mathrm{AMSE} = \mathrm{tr}\{\mathrm{Var}_{\pi} g - \mathrm{Cov}_{\pi}(g,\eta)[\mathrm{Var}_{\pi}\eta + \mathrm{E}_{\pi}(\sigma^2)W]^{-1} \mathrm{Cov}_{\pi}(\eta, g)\}.$$

2. Intrinsic linearization

Besides linear models, estimators in intrinsically linear models still have very good statistical properties. (The model (1) is called intrinsically linear, if its expectation surface

$$\mathcal{E}_\eta = \{\eta(\theta); \theta \in \Theta\} \tag{8}$$

is relatively open set of a s-dimensional plane of $\mathbb{R}^N$, where $s \le m$ (Def. 2.2.1 in [6])). The method of [7] parametrically linearize even intrinsically linear models, which is often not necessary from statistical point of view. Therefore here we present another method which approximates nonlinear model by intrinsically linear one, so that the models which are originally intrinsically linear are not modified.

In the following example we show that linearization by smoothing can indeed change expectation surface of intrinsically linear model very much.

Example 1. Let us consider intrinsically linear model

$$\eta(\theta) = (\cos^2\theta, \sin^2\theta)^\top; \theta \in \Theta \subseteq \mathbb{R}^1.$$

Let prior $\pi(.)$ is proper uniform probability distribution on Θ.

Expectation surface on the $\mathrm{supp}(\pi(.))$ of the model is $\{(t, 1-t)^\top; t \in <0; 1>\}$ in both cases a), b) considered below.

a) let $\Theta = (0; 2\pi)$

Then the linearization by smoothing is singular:

$$A = \bar{0}, \qquad a = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}, \qquad K_1 = \frac{1}{4},$$

so that the expectation surface of linearization by smoothing is

$$\{A\theta + a; \theta \in \Theta\} = \left\{ \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} \right\}.$$

b) let $\Theta = (0; \frac{\pi}{2})$

Then

$$A = \frac{12}{\pi(\pi^2-4)} \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad a = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} - \frac{\pi}{4} A, \quad K_1 = \frac{1}{4} - \frac{12}{\pi^2(\pi^2-4)} \doteq 0.043$$

and the expectation surface of linearization by smoothing is an interval

$$\{A\theta + a; \theta \in \Theta\} = \left\{ \begin{pmatrix} t \\ 1-t \end{pmatrix} : t \in \left(-\frac{3}{\pi^2-4} + \frac{1}{2}; \frac{3}{\pi^2-4} + \frac{1}{2} \right) \doteq (-0.011; 1.011) \right\}$$

In contrast, the best intrinsic linearization problem can be understood as problem of best approximation of the expectation surface of original model (1) by the expectation surface of some intrinsically linear model.

Let

$$y = \xi(\theta) + \varepsilon \, ; \qquad \theta \in \Theta, \tag{9}$$
$$\varepsilon \sim N(0, \sigma^2 W)$$

be an intrinsically linear model. It is known, (see [6]), that for every intrinsically linear model there exists a parametrization, in which the model is a regular linear model. Therefore expectation surface of model (9) is

$$\mathcal{E}_\xi = \mathcal{E}_{A,a} = \left\{ A\beta + a \, : \, \beta \in \mathbb{R}^k \right\},$$

for some reparametrization $\beta = \beta(\theta)$, $A \in \mathbb{R}^{N \times k}$, $\mathrm{rank}(A) = k$, $a \in \mathbb{R}^N$, $k \in \mathbb{N}$. The distance of a point $\eta(\theta) \in \mathcal{E}_\eta$ from $\mathcal{E}_\xi$ in space $\mathbb{R}^N$ with scalar product $< a, b >_W := a^\top W^{-1} b$ is

$$d[\eta(\theta), \mathcal{E}_\xi] := \min_{z \in \mathcal{E}_\xi} \|\eta(\theta) - z\|_W^2,$$

and

$$z_{opt}(\theta) := \arg \min_{z \in \mathcal{E}_{A,a}} \|\eta(\theta) - z\|_W^2$$

is a W-orthogonal projection of $\eta(\theta)$ on $\mathcal{E}_\xi$.

Let $\beta(\theta) \in \mathbb{R}^k$ be such that

$$z_{opt}(\theta) = A\beta(\theta) + a.$$

Then

$$\beta(\theta) = (A^\top W^{-1} A)^{-1} A^\top W^{-1} (\eta(\theta) - a). \tag{10}$$

If we have some prior guess on the plausible values θ in the form of a prior weight function $\pi(\theta)$, the global distance of $\mathcal{E}_\xi$ from $\mathcal{E}_\eta$ can be measured by

$$\begin{aligned} d_\pi(\mathcal{E}_\eta, \mathcal{E}_\xi) : &= \int_\Theta \|\eta(\theta) - z_{opt}(\theta)\|^2_W \pi(\theta) d\theta = \\ &= \int_\Theta \|\eta(\theta) - (A\beta(\theta) + a)\|^2_W \pi(\theta) d\theta = \\ &= \int_\Theta \|\eta(\theta) - [A(A^\top W^{-1} A)^{-1} A^\top W^{-1}(\eta(\theta) - a) + a]\|^2_W \pi(\theta) d\theta. \end{aligned}$$

Let q be the dimension of manifold $\mathcal{E}_\eta$.

In problem of intrinsic linearization of model (1) we consider as optimal such a choice of A and a which is a solution of

$$\min_{\substack{k \leq q \\ A \in \mathbb{R}^{N \times k} \\ \mathrm{rank}(A) = k \\ a \in \mathbb{R}^N}} d_\pi(\mathcal{E}_\eta, \mathcal{E}_{A,a}). \tag{11}$$

The solution of (11) is given in the following statement.

Theorem. *The optimal choice of k, A, a is equal to*

$$k = \min\{q,\ \textit{number of nonzero eigenvalues of } \mathrm{Var}_\pi(W^{-1/2}\eta)\},$$

$$A = W^{1/2}(u_1, \dots, u_k), \tag{12}$$

where $u_1, \dots, u_N$ are orthonormal eigenvectors corresponding to eigenvalues $\lambda_1 \geq \cdots \geq \lambda_N \geq 0$ of the matrix

$$\mathrm{Var}_\pi(W^{-1/2}\eta),$$

respectively,

(12a) $$a \in \mathrm{E}_{\pi}\eta + \mathcal{K}er(I - A(A^{\top}W^{-1}A)^{-1}A^{\top}W^{-1}).$$

There is exactly one $k-$dimensional affine manifold $\mathcal{E}_{\xi}$ such that for every optimal choice of A, a it holds that $\mathcal{E}_{\xi} = \mathcal{E}_{A,a}$.

$$\min_{\substack{k\leq q\\ A\in\mathbb{R}^{N\times k}\\ rank(A)=k\\ a\in\mathbb{R}^{N}}} d_{\pi}(\mathcal{E}_{\eta}, \mathcal{E}_{A,a}) = \sum_{i=k+1}^{N}\lambda_i = \sum_{i=q+1}^{N}\lambda_i.$$

Proof.

$$\begin{aligned}
d_{\pi}(\mathcal{E}_{\eta}, \mathcal{E}_{A,a}) =\ & \mathrm{E}_{\pi}\left\{\left\|(I - A(A^{\top}W^{-1}A)^{-1}A^{\top}W^{-1})(\eta(\theta) - a)\right\|_{W}^{2}\right\}\\
&= tr\left\{(I - A(A^{\top}W^{-1}A)^{-1}A^{\top}W^{-1})(a - \mathrm{E}_{\pi}\eta)\right\}^{\top}\\
&\qquad \cdot W^{-1}\left\{(I - A(A^{\top}W^{-1}A)^{-1}A^{\top}W^{-1})(a - \mathrm{E}_{\pi}\eta)\right\}\\
&\quad +\ tr\ W^{-1}\,\mathrm{Var}_{\pi}\left\{(I - A(A^{\top}W^{-1}A)^{-1}A^{\top}W^{-1})\eta\right\}\\
&\geq tr\ W^{-1}\,\mathrm{Var}_{\pi}\left\{(I - A(A^{\top}W^{-1}A)^{-1}A^{\top}W^{-1})\eta\right\},
\end{aligned}$$

with equality iff $a \in \mathrm{E}_{\pi}\eta + \mathcal{K}er(I - A(A^{\top}W^{-1}A)^{-1}A^{\top}W^{-1})$.

Now it is sufficient to solve the minimization problem

$$\min_{\substack{k\leq q\\ A\in\mathbb{R}^{N\times k}\\ \mathrm{rank}(A)=k}} \mathrm{tr}\ W^{-1}\,\mathrm{Var}_{\pi}\left\{(I - A(A^{\top}W^{-1}A)^{-1}A^{\top}W^{-1})\eta\right\}.$$

Since the matrix $A(A^\top W^{-1}A)^{-1}A^\top W^{-1}$ is idempotent, the last problem is equivalent with the following one:

$$\max_{\substack{k\leq q\\ A\in\mathbb{R}^{N\times k}\\ \mathrm{rank}(A)=k}} \ \mathrm{tr}\ (W^{-1}A(A^\top W^{-1}A)^{-1}A^\top W^{-1}\,\mathrm{Var}_{\pi}\eta)$$

or (since A is optimal solution iff AD is optimal solution for arbitrary regular $D_{k\times k}$)

$$\max_{\substack{k\leq q\\ A\in\mathbb{R}^{N\times k}\\ \mathrm{rank}(A)=k\\ A^\top W^{-1}A=I_k}} \ \mathrm{tr}\ (A^\top W^{-1/2}\,\mathrm{Var}_{\pi}[W^{-1/2}\eta]W^{-1/2}A).$$

The last expression is a problem of principal component analysis of random quantity $W^{-1/2}\eta$, from which it follows that the solution has the form given in the statement of the theorem. □

The obtained intrinsically linear approximation of the original model (1) is equal to

$$\begin{aligned} &y = A(A^\top W^{-1}A)^{-1}A^\top W^{-1}(\eta(\theta) - \mathrm{E}_{\pi}\eta) + \mathrm{E}_{\pi}\eta + \varepsilon,\\ &\varepsilon \sim N(0,\sigma^2 W), \end{aligned} \tag{13}$$

with A taken according to Theorem.

It is obvious that for arbitrary prior π, the original model (1), is intrinsically linear (with $\pi-$probability 1) iff $\mathrm{rank}(\mathrm{Var}_{\pi}\eta) \leq q$.

From Theorem it also follows that the minimal squared “distance”

$$D_1 := \sum_{i=k+1}^{N} \lambda_i \tag{14}$$

of the linearized model (13) from the original model (1) can be understood as measure of intrinsic nonlinearity of model (1).

Example 2 (*continuing example 1)*. In both cases a) and b) the optimal matrices for intrinsic linearization are

$$A = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \quad a = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}, \quad D_1 = 0,$$

so the model (9) has the form

$$y = \begin{pmatrix} \cos^2\theta \\ \sin^2\theta \end{pmatrix} + \varepsilon,$$
$$\varepsilon \sim N(0, \sigma^2 W).$$

Example 3. Let

$$y = \begin{pmatrix} \theta \\ c\theta^2 \end{pmatrix} + \varepsilon, \qquad \theta \in \Theta = \langle -1, 1 \rangle,$$
$$\varepsilon \sim N(0, \sigma^2 I_{2\times 2}),$$

where c is some known positive constant. Let θ has proper uniform prior distribution π on Θ. Then the expectation of linearization by smoothing is

$$(\theta, c/3)^\top.$$

The expectation of intrinsic linearization is

$$\begin{cases} (\theta, c/3)^\top, & \text{if } c < \sqrt{5/3}, \\ (0, c\,\theta^2)^\top, & \text{if } c > \sqrt{5/3}, \end{cases}$$

i.e. in the latter case the expectation surface of intrinsic linearization is orthogonal to that of linearization by smoothing. If $c = \sqrt{5/3}$, the matrix $\mathrm{Var}_{\pi}\eta$ has two identical eigenvalues, so that intrinsic linearization is not uniquely determined.

Example 4. In [8] two three-parameter sigmoidal models are considered for data set 1 from Appendix 4A on a vegetative growth process. The models are

$$\begin{aligned} y_i = \eta(\theta, x_i) + \varepsilon_i = \theta_1 \exp[-\exp(\theta_2 - \theta_3 x_i)] + \varepsilon_i \ , &\qquad i = 1, \dots, N, \\ \varepsilon = (\varepsilon_1, \dots, \varepsilon_N)^\top \sim N(0, \sigma^2 W) &\qquad \text{(Gompertz model)}, \end{aligned}$$

and

$$\begin{aligned} y_i = \eta(\theta, x_i) + \varepsilon_i = \frac{\theta_1}{1 + \exp(\theta_2 - \theta_3 x_i)} + \varepsilon_i, &\qquad i = 1, \dots, N, \\ \varepsilon = (\varepsilon_1, \dots, \varepsilon_N)^\top \sim N(0, \sigma^2 W), &\qquad \text{(logistic model)}, \end{aligned}$$

with σ^2 unknown, $W = I$. We consider here two normal prior distributions – $N(\hat{\theta}_{ML}, s^2(y)M(\hat{\theta}_{ML}))$ (1), and $N(\hat{\theta}_{ML}, 25s^2(y)M(\hat{\theta}_{ML}))$ (2), where $\hat{\theta}_{ML}$ is maximum likelihood estimate in model (1),

$$s^2(y) := \frac{\|y - \eta(\hat{\theta}_{ML})\|_W^2}{N - m},$$

and

$$M(\theta) := \frac{\partial \eta^\top(\theta)}{\partial \theta} W^{-1} \frac{\partial \eta(\theta)}{\partial \theta^\top}.$$

Then the results based on 10000 simulations from prior distribution are (K_{int} and K_{par} are intrinsic and parametric nonlinearity measures from [1]):

Nonlinearity measure	Prior	Gompertz	Logistic
K_{int}		$9.010{\cdot}10^{-2}$	$7.300{\cdot}10^{-2}$
K_{par}		$2.324{\cdot}10^{0}$	$6.440{\cdot}10^{-1}$
K_1	1	$3.000{\cdot}10^{0}$	$1.000{\cdot}10^{-1}$
	2	$1.000{\cdot}10^{3}$	$1.000{\cdot}10^{2}$
D_1	1	$4.000{\cdot}10^{-3}$	$1.000{\cdot}10^{-3}$
	2	$1.500{\cdot}10^{0}$	$4.000{\cdot}10^{-1}$
AMSE	1	$2.000{\cdot}10^{1}$	$1.600{\cdot}10^{0}$
	2	$3.000{\cdot}10^{2}$	$1.500{\cdot}10^{1}$

3. Nonlinear regression inference using intrinsically linear approximation

Merits and shortcomings of the above described linearization methods will be compared at examples of point estimation and construction of confidence regions for parameter θ. However, since computation of ML-estimate in intrinsically linearized model is no easier than in original model, the importance of intrinsic linearization is greater in interval estimation and prediction.

Example 5 (*continuing Example 4*)**.** For data y from [8] and for 10000 simulations from prior distribution we get ML-estimates in original model (i), in linearization by smoothing (ii), in intrinsic linearization (iii), and linear explicit estimate (iv) of parameter θ:

Nonlinearity measure	Prior	Method	Gompertz	Logistic
θ_1	1	(i)	$8.283\cdot10^{1}$	$7.246\cdot10^{1}$
		(ii)	$8.360\cdot10^{1}$	$7.254\cdot10^{1}$
		(iii)	$8.290\cdot10^{1}$	$7.245\cdot10^{1}$
		(iv)	$8.300\cdot10^{1}$	$7.249\cdot10^{1}$
	2	(ii)	$1.100\cdot10^{2}$	$7.500\cdot10^{1}$
		(iii)	$8.400\cdot10^{1}$	$7.260\cdot10^{1}$
		(iv)	$8.700\cdot10^{1}$	$7.250\cdot10^{1}$
θ_2	1	(i)	$1.224\cdot10^{0}$	$2.618\cdot10^{0}$
		(ii)	$1.230\cdot10^{0}$	$2.623\cdot10^{0}$
		(iii)	$1.223\cdot10^{0}$	$2.619\cdot10^{0}$
		(iv)	$1.227\cdot10^{0}$	$2.621\cdot10^{0}$
	2	(ii)	$1.400\cdot10^{0}$	$2.720\cdot10^{0}$
		(iii)	$1.200\cdot10^{0}$	$2.610\cdot10^{0}$
		(iv)	$1.300\cdot10^{0}$	$2.660\cdot10^{0}$
θ_3	1	(i)	$3.710\cdot10^{-2}$	$6.740\cdot10^{-2}$
		(ii)	$3.720\cdot10^{-2}$	$6.750\cdot10^{-2}$
		(iii)	$3.700\cdot10^{-2}$	$6.740\cdot10^{-2}$
		(iv)	$3.710\cdot10^{-2}$	$6.730\cdot10^{-2}$
	2	(ii)	$3.400\cdot10^{-2}$	$6.940\cdot10^{-2}$
		(iii)	$3.600\cdot10^{-2}$	$6.720\cdot10^{-2}$
		(iv)	$3.600\cdot10^{-2}$	$6.830\cdot10^{-2}$

There are several kinds of $(1-\alpha)$-confidence regions for parameter θ used in nonlinear regression (see discussion in [5]). Here we compare the following ones:

regions based on likelihood ratio (exact only in intrinsically linear models)

$$\Theta_{LR} := \left\{\theta \in \Theta; \frac{(N-m)(\|y-\eta(\theta)\|_W^2 - \|y-\eta(\hat{\theta})\|_W^2)}{m\|y-\eta(\hat{\theta})\|_W^2} \le F_{m,M-m}(1-\alpha)\right\}, \tag{15}$$

where $\hat{\theta}$ is ML-estimate in model (1),

and regions based on projections

$$\Theta_P := \left\{ \theta \in \Theta; \frac{(N-p)\|P(y-\eta(\theta))\|_W^2}{p\|(I-P)(y-\eta(\theta))\|_W^2} \le F_{p,N-p}(1-\alpha) \right\} \tag{16}$$

where P is $W-$orthogonal projector, $p = \text{rank}(P)$. If P does not depend on y, region (16) is exact in arbitrary model.

Note, It can happen that no region of the form (16) is good, since the set $\mathcal{M}_P$ of points $\mu \in \mathbb{R}^N$ which satisfy definition inequality of region (16) is a cone (differently from the set $\mathcal{M}_{LR}$ which is a ball), and intersections of expectation surfaces of some models with such cones may be unbounded, too large, or, on the contrary, void sets. See Figure 6.19 in [2].

For above given linearization methods it is natural to construct confidence regions for θ of the form (16) with

$$P = A(A^\top W^{-1} A)^{-1} A^\top W^{-1},$$

where A is optimal matrix from (2) (with $\theta^0 = \hat{\theta}_{ML}$), (5), (7) (with $g(\theta) := \eta(\theta)$) or (12). Corresponding P will be denoted P_{ML}, P_{SM}, P_{EX}, P_{IN}, respectively. (Region for P_{ML} without parts due to overlapping is almost exact in flat models (see [5]). Among this class of confidence regions the confidence regions based on intrinsic linearization, i.e. on projector P_{IN} should violate the objection from the Note against confidence regions of type (16) in the most vigorous degree possible since $(I-P)$ in this case corresponds to the "direction" of apriori shortest diameter of expectation surface. If the prior used is subjective, then intersection of such confidence region with the support of prior can be used.

Example 6 (*continuing Example 3*)**.** Point estimates of θ:
Linearization by smoothing: $\hat{\theta} = y_1$.

Intrinsic linearization:

$$\hat{\theta} = \begin{cases} y_1, & \text{if } c < \sqrt{5/3}, \\ \sqrt{y_2/c}, & \text{if } c > \sqrt{5/3} \quad \text{and} \quad y_2 \geq 0, \\ 0, & \text{if } c > \sqrt{5/3} \quad \text{and} \quad y_2 < 0. \end{cases}$$

Linear explicit estimation: $\hat{\theta} = 1/(1 + 3\,\mathrm{E}_{\pi_2}(\sigma^2))\, y_1 < y_1$, where π_2 is a prior distribution for σ^2, which is assumed to be independent of θ.

Since the prior π is uniform, ML-estimator of θ equals to posterior modus estimator justified from the bayesian point of view. Therefore, quality of estimators from various linearizations can be assessed by their closeness to ML-estimator. Expressions for estimators of θ give an idea which linearization is suitable for different values of y. It can be roughly recommended to use

linearization by smoothing,	if y_2 is small and y is above the parabola,
linear explicit estimation,	if y_2 is small and y is under the parabola,
intrinsic linearization,	if y_2 is large.

Let us consider the case favorable to intrinsic linearization with $c = 10$, $\mathrm{E}_{\pi_2}(\sigma^2) = 0.3^2$, and $y = (0.87; 8.4)^\top$. Then we get point estimates:

$$\hat{\theta} = \begin{cases} 9.16377 \cdot 10^{-1}, & \text{(ML-estimate)}, \\ 8.70000 \cdot 10^{-1}, & \text{(linearization by smoothing)}, \\ 9.16515 \cdot 10^{-1}, & \text{(intrinsic linearization)}, \\ 6.85000 \cdot 10^{-1}, & \text{(linear explicit estimation)} \end{cases}$$

and 0.9-confidence intervals for θ

$\langle 9.005 \cdot 10^{-1}\,;\, 9.320 \cdot 10^{-1}\rangle$,

LR conf. region,

$\langle -1.000 \cdot 10^{1}\,;\, -4.669 \cdot 10^{-1}\rangle \cup \langle 9.003 \cdot 10^{-1}\,;\, 9.321 \cdot 10^{-1}\rangle$,

LR conf. region (16) with $P = P_{ML}$,

$\langle -1.000 \cdot 10^{1}\,;\, -9.320 \cdot 10^{-1}\rangle \cup \langle -9.010 \cdot 10^{-1}\,;\, 9.161 \cdot 10^{-1}\rangle \cup \langle 9.169 \cdot 10^{-1}\,;\, 1.000 \cdot 10^{1}\rangle$,

conf. region (16) with $P = P_{SM} = P_{EX}$,

$\langle -1.000 \cdot 10^{1}\,;\, -3.095 \cdot 10^{-1}\rangle \cup \langle 9.046 \cdot 10^{-1}\,;\, 9.404 \cdot 10^{-1}\rangle$,

conf. region (16)with $P = P_{IN}$.

First parts of regions for $P = P_{ML}$ and $P = P_{IN}$ and first and third part of region for $P = P_{SM} = P_{EX}$ are due to overlapping.

Acknowledgment. I am grateful to Professor Andrej Pázman for discussion on the subject of this paper.

1. Bates D. M. and Watts, D. G., *Relative curvature measures of nonlinearity*, J. R. Statist. Soc. B **42** (1980), 1–25.
2. ______, *Nonlinear regression analysis and its applications*, J. Wiley, New York 1988.
3. Gallant A. R., *Explicit estimators of parametric functions in nonlinear regression*, J. Am. Stat. Assoc. **75** 369 (1980), 182–193.
4. Kubáček L., *On a linearization of regression models*, Appl. Math. **40** 1 (1995), 61–78.
5. Pázman A., *A classification of nonlinear regression models and parameter confidence regions*, Kybernetika **28** (1992), 444–453.
6. ______, *Nonlinear statistical models*, Kluwer, Dordrecht 1993.
7. ______, *Linearization of nonlinear regression models by smoothing*, Tatra Mt. Math. Publ. **22** (2001), 13–25.
8. Ratkowsky D. A., *Nonlinear regression modeling* M. Dekker, New York 1983.

K. Hornišová, Institute of Measurement Science, Slovak Academy of Sciences Bratislava, Dúbravská 9, 841 04 Bratislava, *e-mail*: `umerhorn@savba.sk`