Acta Math. Univ. Comenianae
Vol. LXI, 2(1992), pp. 263–276 

# GEOMETRY OF THE NONLINEAR REGRESSION WITH PRIOR

A. PÁZMAN

Abstract. In a nonlinear regression model with a given prior distribution, the estimator maximizing the posterior probability density is considered (a certain kind of Bayes estimator). It is shown that the prior influences essentially, but in a comprehensive way, the geometry of the model, including the intrinsic curvature measure of nonlinearity which is derived in the paper. The obtained geometrical results are used to present the modified Gauss-Newton method of computation of the estimator, and to obtain the exact and an approximate probability density of the estimator.

## 1. Introduction

We consider the nonlinear regression model

$$\begin{aligned} &y = \eta(\vartheta) + \varepsilon; \quad (\vartheta \in \Theta) \\ &\varepsilon \sim N(0, \Sigma)\,, \end{aligned} \tag{1.1}$$

with the observed vector $y \in R^N$, a closed parameter space $\Theta \subseteq R^m$, $m < N$, $\operatorname{int}(\Theta) \neq \emptyset$. We suppose that the mapping $\eta(\cdot)$ is one-to-one and continuous on $\Theta$, has continuous second order derivatives on $\operatorname{int}(\Theta)$, and that the matrix

$$J(\vartheta) \, : \, = \nabla_\vartheta^T \eta(\vartheta)$$

has a full rank for every $\vartheta \in \operatorname{int}(\Theta)$. The error variance matrix $\Sigma$ is supposed to be known, and positive definite.

We note, that the assumption of normality of the error vector $\varepsilon$ is not needed in the geometrical and numerical considerations in Sections 2 and 3.

Further we suppose that a prior density $\pi(\vartheta)$ is given, and we shall suppose that the function $\pi(\cdot)$ has continuous second order derivatives and that its support,

Received April 6, 1992.
1980 *Mathematics Subject Classification* (1991 *Revision*). Primary 62J02; Secondary 62F15, 62F11.
*Key words and phrases.* nonlinear regression, Bayes estimator, distributions of estimators, geometry in statistics, curvatures.

supp$\pi \in \mathrm{int}(\Theta)$. If the prior is given, any meaningful estimator of $\vartheta$ must take it into account. We propose to use the estimator

$$\hat{\vartheta} := \hat{\vartheta}(y) := \arg\max_{\vartheta \in \Theta} \pi(\vartheta|y)\,, \tag{1.2}$$

where $\pi(\vartheta|y)$ is the posterior probability density of $\vartheta$. Hence $\hat{\vartheta}$ is the modus of the posterior density. Using the Bayes formula for the posterior density, we can write

$$\begin{aligned}\hat{\vartheta} &= \arg\max_{\vartheta \in \Theta} \pi(\vartheta) f(y|\vartheta) \\ &= \arg\min_{\vartheta \in \Theta} Z(\vartheta, y)\,,\end{aligned}$$

where

$$Z(\vartheta, y) := \|y - \eta(\vartheta)\|_{\Sigma}^2 - 2l(\vartheta)\,.$$

Here $f(y|\vartheta)$ is the normal density of $y$, given $\vartheta$, and $l(\vartheta) := \ln \circ \pi(\vartheta)$. We use the notation $\langle a, b\rangle_{\Sigma} := a^T \Sigma^{-1} b$, $\|a\|_{\Sigma} := [a^T \Sigma^{-1} a]^{1/2}$, $\|a\| := \|a\|_I$.

We have several justifications for the use of this estimator.

a) Suppose that $l(\cdot)$ is zero on some set $\Theta^* \subseteq \mathrm{int}(\Theta)$, and is decreasing smoothly to minus infinity when $\vartheta$ is approaching to the boundary of $\Theta$. Such a choice of $l(.)$ allows to express quantitatively that the boundaries of $\Theta$ are uncertain, and to maintain the maximum likelihood estimator unchanged on $\Theta^*$. This case is considered in [**7**] (without considering the curvature), and used there for experimental design.

b) If $\pi(\vartheta)$ is the likelihood function obtained from some previous (independent) experiment, then $\hat{\vartheta}$ is simply the maximum likelihood estimator obtained from both, the previous and the actual experiments. The results presented in Section 4 allow to investigate the influence of the distribution of the observed vector in the actual experiment, on the distribution of the estimator $\hat{\vartheta}$ obtained from both experiments. This can be useful e.g. in batch-sequential design of experiments (cf. [**3**]).

c) In the case of a linear model, $\eta(\vartheta) = F\vartheta$, with a normal prior

$$\pi(\vartheta) = (2\pi)^{-m/2} \det{}^{-1/2}(C) \exp\left\{-\frac{1}{2}\|\vartheta\|_C^2/2\right\}\,,$$

we obtain by a direct computation that

$$\vartheta = (F^T \Sigma^{-1} F + C^{-1})^{-1} F^T \Sigma^{-1} y$$

which is the standard Bayes estimator in linear models (i.e. it is the mean of $\pi(\vartheta|y)$). However, in nonlinear models the mean of $\pi(\vartheta|y)$ is not equal to the modus of $\pi(\vartheta|y)$. We think that the posterior modus estimator should be preferred in this case, as being closely related to the maximum likelihood estimator standardly used in nonlinear models (without prior).

When the maximum likelihood estimator is used, well known geometric objects are:

a) The expectation surface

$$\mathscr{E} : = \{\eta(\vartheta) : \vartheta \in \Theta\}.$$

b) The tangent plane at $\vartheta$

$$\mathscr{T}_\vartheta \; : = \{\eta(\vartheta) + J(\vartheta)v; \; v \in R^m\}.$$

c) The ancillary space for the maximum likelihood estimator

$$\mathscr{A}_\vartheta \; : = \{y \in R^N : \; [y - \eta(\vartheta)]^T \Sigma^{-1} J(\vartheta) = 0\}.$$

d) The intrinsic curvature of Bates and Watts (cf. [**2**], [**4**])

$$K(\vartheta) \; : = \sup_{v \in R^m, \|v\|=1} \frac{\|[I - P(\vartheta)][v^T H(\vartheta) v]\|_\Sigma}{v^T M(\vartheta) v}. \tag{1.3}$$

e) The parameter effect curvature, which is given by a formula similar to (1.3).

**Notation.** In (1.3) we denote by $M(\vartheta)$ the Fisher information matrix

$$M(\vartheta) \; : = J^T(\vartheta)\Sigma^{-1}J(\vartheta).$$

The matrix

$$P(\vartheta) \; : = J(\vartheta)M^{-1}(\vartheta)J^T(\vartheta)\Sigma^{-1}$$

is the $\langle\ ,\ \rangle_\Sigma$-orthogonal projector onto the tangent space

$$\mathscr{L}_\vartheta \; : = \{J(\vartheta)v : v \in R^m\}$$

which is parallel to $\mathscr{T}_\vartheta$. $H(\vartheta)$ is a 3-dimensional array with entries

$$\{H(\vartheta)\}_{ij}^k = \partial^2 \eta_k(\vartheta)/\partial\vartheta_i \partial\vartheta_j.$$

The multiplication of $H(\vartheta)$ with the matrix $(I - P(\vartheta))$ is taken over the superscript $k$, and the multiplications with $v^T$ and with $v$ are over the subscripts $i$ and $j$. In general, the dimension of the multiplied vector or matrix shows which subscript or superscript of $H(\vartheta)$ is to be used. In (1.3) we have

$$v^T H v = \sum_{i,j} v_i \frac{\partial^2 \eta(\vartheta)}{\partial\vartheta_i \partial\vartheta_j} v_j.$$

This vector is multiplied by the matrix $I - P(\vartheta)$, and then its norm $\|\ \|_\Sigma$ is taken.

In the paper we consider the influence of the prior on the geometry. In particular we derive a new expression for the curvature (see eq. (2.4) and Theorem 1), and we use the geometric interpretation of the model for the numerical computation of estimates and for the computation of the probability density of the estimator (Theorems 2 and 3).

## 2. The Changes in The Geometry of The Model Due To The Prior

The "normal equation" related to the estimator $\hat{\vartheta}$ has the form

$$\nabla_\vartheta Z(\vartheta, y) = 0$$

i.e.

$$J^T(\vartheta)\Sigma^{-1}[\eta(\vartheta) - y] - \nabla_\vartheta l(\vartheta) = 0\,. \tag{2.1}$$

Denote by $u(\vartheta)$ the vector

$$u(\vartheta) := J(\vartheta)M^{-1}(\vartheta)\nabla_\vartheta l(\vartheta)\,.$$

Let us multiply (2.1) by the matrix $J(\vartheta)M^{-1}(\vartheta)$. We obtain

$$P(\vartheta)[\eta(\vartheta) - y] = u(\vartheta)\,, \tag{2.2}$$

or equivalently

$$P(\vartheta)[\eta(\vartheta) - u(\vartheta) - y] = 0\,.$$

Reversely, multiplying (2.2) from the left by the matrix $J^T(\vartheta)\Sigma^{-1}$, we obtain (2.1). Hence (2.2) is another form of the normal equation for $\hat{\vartheta}$. In comparison, the normal equation for the maximum likelihood estimator is equal to

$$P(\vartheta)[\eta(\vartheta) - y] = 0\,, \tag{2.3}$$

i.e. the estimator is simply a projector of $y$ onto $\mathscr{E}$. In (2.2) we have besides the projection also a shift by $u(\vartheta)$.

According to [**1**], the ancillary space of an estimator is the set of all samples giving the same solution $\vartheta$ of the normal equation. The ancillary space corresponding to the estimator $\hat{\vartheta}$ is equal to

$$\mathscr{B}_{\hat{\vartheta}} := \{y \in R^N : P(\hat{\vartheta})[\eta(\hat{\vartheta}) - y] = u(\hat{\vartheta})\}\,.$$

The ancillary space of the maximum likelihood estimator is obtained when $u(\hat{\vartheta}) = 0$, and it is equal to the set $\mathscr{A}_{\hat{\vartheta}}$ given in Section 1. Evidently, the plane $\mathscr{B}_\vartheta$ is parallel to $\mathscr{A}_\vartheta$, and

$$\mathscr{B}_\vartheta = \mathscr{A}_\vartheta + \{u(\vartheta)\}\,.$$

**Lemma 1.** *The "shift vector" $u(\vartheta)$ is invariant with respect to any regular change of parameters in the regression model. It is orthogonal to both planes $\mathscr{A}_\vartheta$ and $\mathscr{B}_\vartheta$. Its "length" is equal to*

$$\|u(\vartheta)\|_\Sigma = [\nabla_\vartheta^T l(\vartheta) M^{-1}(\vartheta) \nabla_\vartheta l(\vartheta)]^{1/2} .$$

*Proof.* The proof is obvious. □

Now we consider the "intrinsic" curvature of model (1.1). When the estimator (1.2) is used, we propose to express the intrinsic curvature of the model by the expression

$$K_\pi(\vartheta) \ : = \sup_{v\in R^m, \|v\|=1} \frac{\|[I-P(\vartheta)][v^T H(\vartheta) v]\|_\Sigma}{v^T[M(\vartheta)+G(\vartheta)]v}\,, \tag{2.4}$$

where

$$G(\vartheta) \ : = \ -\nabla\nabla^T l(\vartheta) + u^T(\vartheta)\Sigma^{-1}H(\vartheta)\,.$$

We note that according to the notation convention presented after eq. (1.3), we denote by $u^T(\vartheta)\Sigma^{-1}H(\vartheta)$ an $m\times m$ matrix with entries

$$\sum_{k=1}^N \{\Sigma^{-1}u(\vartheta)\}_k \frac{\partial^2 \eta_k(\vartheta)}{\partial\vartheta_i\partial\vartheta_j}\,; \quad (i,j=1,\dots,m)\,.$$

The expression (2.4) makes sense only when the matrix $M(\vartheta)+G(\vartheta)$ is positive semi definite, which gives some (natural) restriction on the considered prior $\pi(\vartheta)$. This will be discussed later. We note that $K_\pi(\vartheta)=\infty$ if $M(\vartheta)+G(\vartheta)$ is positive semi definite but not definite.

We have $K_\pi(\vartheta)=0$ when the model is linear. In the case that $\pi(\vartheta)=$ const., i.e. when there is no prior information, we have $G(\vartheta)=0$, and $K_\pi(\vartheta)$ coincides with (1.3).

To derive (to justify) the formula (2.4) we use the following geometrical approach: We take the ancillary space $\mathscr{B}(\vartheta)$, and we consider the "neighbor" ancillary space $\mathscr{B}(\vartheta+\delta)$ for different small $\delta\in R^m$. If $y\in\mathscr{B}(\vartheta)\cap\mathscr{B}(\vartheta+\delta)$, then the solution of the normal equation is ambiguous. To avoid such ambiguities (for any sufficiently small $\delta$) the distance of $y\in\mathscr{B}(\vartheta)$ from the $\mathscr{T}_\vartheta$ (i.e. the value of $\|(I-P(\vartheta))(y-\eta(\vartheta))\|_\Sigma$) should not be too large. It is therefore statistically and geometrically meaningful to take as the effective radius of curvature at the point $\vartheta$ (denoted by $\rho_\pi(\vartheta)$) the smallest distance of the set

$$\bigcup_{\delta\in R^m, \|\delta\|\le\Delta} \mathscr{B}(\vartheta)\cap\mathscr{B}(\vartheta+\delta)$$

from the set $\mathscr{T}_\vartheta$, when $\Delta$ tends to zero. We can write

$$\begin{aligned}
\bigcup_{\delta\in R^m,\|\delta\|\le\Delta} & \mathscr{B}(\vartheta)\cap\mathscr{B}(\vartheta+\delta) \\
&= \bigcup_{\delta\in R^m,\|\delta\|\le\Delta} \{y:\ \nabla_\vartheta Z(\vartheta,y)=0\}\cap\{y:\nabla_\vartheta Z(\vartheta+\delta,y)=0\} \\
&= \bigcup_{\delta\in R^m,\|\delta\|\le\Delta} \{y:\ \nabla_\vartheta Z(\vartheta,y)=0\ \&\ \nabla_\vartheta Z(\vartheta,y) \\
&\qquad + \nabla_\vartheta\nabla_\vartheta^T Z(\vartheta,y)\delta\ + o(\Delta)=0\} \\
&= \{y:\ \nabla_\vartheta Z(\vartheta,y)=0\ \&\ \exists_{\delta\in R^m,\|\delta\|\le\Delta}\ \nabla_\vartheta\nabla_\vartheta^T Z(\vartheta,y)\delta+o(\Delta)=0\}\,.
\end{aligned}$$

Evidently,

$$\begin{aligned}
&\text{there is a } \delta\neq 0 \text{ such that } \nabla_\vartheta\nabla_\vartheta^T Z(\vartheta,y)\delta=0 \\
&\qquad\Leftrightarrow\ \text{the matrix } \nabla_\vartheta\nabla_\vartheta^T Z(\vartheta,y) \text{ is singular} \\
&\qquad\Leftrightarrow \det[\nabla_\vartheta\nabla_\vartheta^T Z(\vartheta,y)]=0\,.
\end{aligned}$$

Correspondingly, we define the effective radius of curvature by the formula

$$\begin{aligned}
\rho_\pi(\vartheta)\ :=\ &\inf\{\|(I-P(\vartheta))(y-\eta(\vartheta))\|_\Sigma:\ y\in R^N\ \&\ \nabla_\vartheta Z(\vartheta,y)=0 \\
&\&\ \det[\nabla_\vartheta\nabla_\vartheta^T Z(\vartheta,y)]=0\}
\end{aligned}$$

**Lemma 2.** *If the matrix $M(\vartheta)+G(\vartheta)$ is positive definite, $y\in\mathscr{B}(\vartheta)$, and*

$$\|(I-P(\vartheta))(y-\eta(\vartheta))\|_\Sigma\ <[K_\pi(\vartheta)]^{-1}$$

*then the matrix $\nabla_\vartheta\nabla_\vartheta^T Z(\vartheta,y)$ is positive definite.*

*Proof.* Since $\vartheta$ is fixed, we shall omit the symbol $\vartheta$ in the abbreviated notation used in the proof.

For any $v\in R^m\setminus\{0\}$ we have

$$v^T(\nabla\nabla^T Z(y))v=(\eta-y)^T\Sigma^{-1}(v^T H v)+v^T M v-v^T(\nabla\nabla^T l)v\,.$$

Using that $P(\eta-y)=u$, one can obtain

$$v^T(\nabla\nabla^T Z(y))v=\big\langle(I-P)(\eta-y),(I-P)(v^T H v)\big\rangle_\Sigma+v^T(M+G)v\,. \tag{2.5}$$

Hence from the Schwarz inequality we obtain

$$\begin{aligned}
v^T(\nabla\nabla^T Z(y))v &\ge [-\|(I-P)(\eta-y)\|_\Sigma\frac{\|(I-P)(v^T H v)\|_\Sigma}{v^T(M+G)v}+1]v^T(M+G)v \\
&\ge 0\,.
\end{aligned}$$

□

**Theorem 1.** *If $\vartheta \in \mathrm{int}\Theta)$ and if the matrix $M(\vartheta) + G(\vartheta)$ is positive definite, then*

$$\rho_\pi(\vartheta) = [K_\pi(\vartheta)]^{-1}\,.$$

*Proof.* From Lemma 2 we obtain

$$\nabla_\vartheta Z(\vartheta, y) = 0 \ \&\ \det[\nabla_\vartheta \nabla_\vartheta^T Z(\vartheta, y)] = 0 \Rightarrow$$
$$\|(I - P(\vartheta))(y - \eta(\vartheta))\|_\Sigma \geq [K_\pi(\vartheta)]^{-1}\,.$$

Hence $\rho_\pi(\vartheta) \geq [K_\pi(\vartheta)]^{-1}$. To prove the inverse inequality, take $v^* \in R^m$, $\|v^*\| = 1$ such that the supremum in (2.4) is attained at $v^*$. Define a point $y^* \in R^m$ by the equalities

i) $P(\vartheta)[\eta(\vartheta) - y^*] = u(\vartheta)$,

ii) $[I - P(\vartheta)][\eta(\vartheta) - y^*] = \lambda(\vartheta)[I - P(\vartheta)](v^{*T} H(\vartheta) v^*)$,

where

$$\lambda(\vartheta) \ :=\ -[K_\pi\ (\vartheta)]^{-1} \|[I - P(\vartheta)](v^{*T} H(\vartheta) v^*)\|_\Sigma^{-1}\,.$$

We can write, like in the proof of Lemma 2

$$\begin{aligned} v^{*T}(\nabla\nabla^T Z(y^*))v^* &= (\eta - y^*)^T \Sigma^{-1}(v^{*T} H v^*) + v^{*T} M v^* - v^{*T}(\nabla\nabla^T l)v^* \\ &= \big\langle (I - P)(\eta - y^*), (I - P)(v^{*T} H v^*)\big\rangle_\Sigma + v^{*T}(M + G)v^* \\ &= 0\,. \end{aligned}$$

Hence $\det[\nabla_\vartheta \nabla_\vartheta^T Z(\vartheta, y^*)] = 0$, and also $\nabla_\vartheta Z(\vartheta, y^*) = 0$, as follows from i). Moreover, from ii) we obtain

$$\|[I - P(\vartheta)][\eta(\vartheta) - y^*]\|_\Sigma = [K_\pi\ (\vartheta)]^{-1}\,.$$

Consequently, from the definition of $\rho_\pi(\vartheta)$ it follows that

$$\rho_\pi(\vartheta) \leq [K_\pi(\vartheta)]^{-1}\,.$$

□

Now we shall show that $K_\pi(\vartheta)$ is an "intrinsic" expression.

**Lemma 3.** *The curvature $K_\pi(\vartheta)$ is invariant to any regular change of parameters in the regression model.*

*Proof.* Let $\beta = \beta(\vartheta)$ be a regular reparameterization of the regression model (i.e. $\det[\nabla_\vartheta \beta(\vartheta)] \neq 0$, and $\nabla\nabla^T \beta(\vartheta)$ is continuous for every $\vartheta \in \mathrm{int}(\Theta)$). Let $\vartheta(\cdot)$ be the mapping inverse to $\beta(\cdot)$. The reparameterized regression model is

$$\begin{aligned} &y = \nu(\beta) + \varepsilon; \quad (\beta \in \beta(\Theta)) \\ &\varepsilon \sim N(0, \Sigma)\,, \end{aligned} \tag{2.6}$$

with $\nu(\beta) := \eta(\vartheta(\beta))$. Let us denote by $P(\beta)$, $M(\beta)$, $G(\beta)$ the matrices corresponding to the model (2.6). We have $I - P(\vartheta) = I - P(\beta)$, since a projector is invariant. Further, by a direct computation of derivatives we obtain

$$(I - P(\beta))\nabla_\beta \nabla_\beta^T \nu(\beta) = \nabla_\beta \vartheta^T(\beta)[(I - P(\vartheta))\nabla_\vartheta \nabla_\vartheta^T \eta(\vartheta)]\nabla_\beta^T \vartheta(\beta)$$

$$M(\beta) = \nabla_\beta \vartheta^T(\beta) M(\vartheta) \nabla_\beta^T \vartheta(\beta)\,,$$

and using the equality (2.2) we obtain

$$\begin{aligned} G(\beta) &= \nabla_\beta \nabla_\beta^T l(\beta) - u^T(\beta)\Sigma^{-1}\nabla_\beta \nabla_\beta^T \nu(\beta) \\ &= \nabla_\beta \vartheta^T(\beta) G(\vartheta) \nabla_\beta^T \vartheta(\beta)\,. \end{aligned}$$

We put these equalities into (2.4), and we use that $\nabla_\beta \vartheta^T(\beta)$ is a regular matrix, to obtain

$$K_\pi(\vartheta) = K_\pi(\beta)\,.$$

□

An interesting interpretation of the curvature $K_\pi(\vartheta)$ can be obtained in terms of geodesic curves.

Let $h\colon t \in (-\delta, \delta) \to h(t) \in \operatorname{int}(\Theta)$ be a curve in $\Theta$ such that $h(\cdot)$ has continuous first and second order derivatives. Let $\gamma(t) := \eta \circ h(t)$ be the corresponding curve on the expectation surface $\mathscr{E}$. The curve $\gamma$ is called a geodesics on $\mathscr{E}$ (and correspondingly $h$ is a geodesics in $\Theta$) iff

i) $\|d\gamma/dt\|_\Sigma = 1$ for every $t$ (i.e. $t$ is the arc-length of the curve $\gamma$)

ii) $\langle d^2\gamma/dt^2, \partial\eta(\vartheta)/\partial\vartheta_i\rangle_\Sigma = 0$; $(i = 1, \dots, m)$ for every $t$ i.e. the curvature vector $d^2\gamma/dt^2$ of the curve $\gamma$ is orthogonal to the expectation surface at every point of the curve $\gamma$; this means that from all curves on $\mathscr{E}$, the geodesic curves are the less curved.

**Lemma 4.** *We can write*

$$K_\pi(\vartheta) = \sup_h \frac{\|d^2\eta \circ h(t)/dt^2\|_\Sigma}{1 - d^2 l \circ h(t)/dt^2}\Big|_{t=0}\,, \tag{2.7}$$

*where the supremum is taken over all geodesics $h$ in $\Theta$ such that $h(0) = \vartheta$.*

*Proof.* Take a geodesics $h$ such that $h(0) = \vartheta$, $[dh(t)/dt]_{t=0} = v$. From the property i) we obtain

$$v^T M(\vartheta) v = 1\,.$$

From the property ii) we have

$$\begin{aligned} d^2\eta \circ h(t)/dt^2 &= (I - P(\vartheta))(d^2\eta \circ h(t)/dt^2) \\ &= (I - P(\vartheta))[v^T H(\vartheta) v]\,. \end{aligned}$$

Performing the derivatives in ii) we obtain

$$M(\vartheta)\frac{d^2h}{dt^2} + J^T(\vartheta)\Sigma^{-1}[v^T H(\vartheta)v] = 0\,,$$

hence

$$\frac{d^2h}{dt^2} = -M^{-1}(\vartheta)J^T(\vartheta)\Sigma^{-1}[v^T H(\vartheta)v]\,.$$

This allows to show that

$$d^2 l \circ h(t)/dt^2|_{t=0} = \text{ - } v^T G(\vartheta)v\,.$$

Thus the expressions (2.4) and (2.7) can be compared directly. □

**Discussion about the case that $M(\vartheta) + G(\vartheta)$ is not positive semi definite:** Take $v \in R^m$, $\|v\| = 1$ such that $v^T(M(\vartheta) + G(\vartheta))v < 0$. Take the geodesic curve $h(.)$ such that $v = dh(t)/dt|_{t=0}$. We have

$$v^T(M(\vartheta) + G(\vartheta))v = 1 - d^2 l \circ h(t)/dt^2|_{t=0}\,.$$

Hence $d^2 l \circ h(t)/dt^2 > 1$ for $t$ in some neighborhood of the point $t = 0$. The equation $d^2 l \circ h(t)/dt^2 = 1$ gives

$$\pi \circ h(t) = \pi \circ h(0)\exp\{\ t^2/2 + \mathrm{ct}\}$$

for some constant $c$, which means an extremely high increase of $\pi \circ h(t)$ in the neighbor of $t = 0$. To avoid cases when $M(\vartheta) + G(\vartheta)$ is not positive semi definite means to avoid a prior density with such extreme local increases.

## 3. The Iterative Computation of Estimates

The geometric ideas presented above help to construct iterative methods for the computation of estimates. We shall show this on a procedure, which is a generalization of the Gauss-Newton method.

In general, in an iterative procedure we chose the point $\vartheta_{n+1}$ according to a rule

$$\vartheta_{n+1} = \vartheta_n + \lambda_n v(\vartheta_n)\,.$$

Here $v(\vartheta_n)$ is the direction of the $n$-th step and $\lambda_n$ is the length of the step. In the case of the maximum likelihood estimator a standard method is the Gauss-Newton method. The geometrical idea of this method is very simple: We project the sample point $y$ onto the tangent plane $T_{\vartheta_n}$. We denote by $\vartheta^*_{n+1}$ the parameter of the obtained point in $T_{\vartheta_n}$. In symbols

$$J(\vartheta_n)(\vartheta^*_{n+1} - \vartheta_n) = P(\vartheta_n)(y - \eta(\vartheta_n))\,, \tag{3.1}$$

and we take the direction of the $n$-th step equal to $\vartheta_{n+1}^* - \vartheta_n$. The step-length is taken either equal to one or another number between 0 and 1, so to ensure the convergence of the procedure.

In the case that we have the prior $\pi(\vartheta)$ we propose to modify (3.1) in the spirit of the geometry presented in Section 2. To the projector $P(\vartheta_n)$ in (3.1) we have to add the shift by $u(\vartheta_n)$. The direction of the $n$-th step $v(\vartheta_n)$ is therefore given by the equation

$$J(\vartheta_n)v(\vartheta_n) = P(\vartheta_n)[y - \eta(\vartheta_n)) + u(\vartheta_n)] .$$

After a multiplication by $J^T(\vartheta_n)\Sigma^{-1}$ we obtain

$$v(\vartheta_n) = M^{-1}(\vartheta_n)[J^T(\vartheta_n)\Sigma^{-1}(y - \eta(\vartheta_n)) + \nabla_\vartheta l(\vartheta_n)] . \tag{3.2}$$

We take the step-length according to

$$\lambda_n \, : = \, \arg \min_{\lambda \in [0,1]} Z(\vartheta_n + \lambda v(\vartheta_n), y) . \tag{3.3}$$

**Theorem 2.** *Let us suppose, that $\pi(\vartheta)$ is bounded, and that* $\mathrm{supp}\pi$ *is a bounded convex set. Take $\vartheta_1$ arbitrary, but $\vartheta_1 \in \mathrm{supp}\pi$. Then*

i) *There is a solution $\vartheta^*$ of* (2.1) *such that*

$$\lim_{n \to \infty} Z(\vartheta_n, y) = Z(\vartheta^*, y) .$$

ii) $\lim_{n\to\infty} a(\vartheta_n) = 0$ , *where $a(\vartheta)$ is the left-hand side of* (2.1).

iii) *The sequence $\vartheta_n$; $n = 1, 2, \dots$ has limit points, and every limit point is a solution of the normal equation* (2.1)

iv) $\lim_{n\to\infty}(\vartheta_{n+1} - \vartheta_n) = 0$ .

*Proof.* We have $-l(\vartheta) = \infty$ (i.e. $Z(\vartheta, y) = \infty$) outside the set $\mathrm{supp}\pi$, hence for every $y$ we have $\hat{\vartheta}(y) \in \mathrm{supp}\pi$, and the assumption $\vartheta_1 \in \mathrm{supp}\pi$ is not restrictive. The set $\mathrm{supp}\pi$ is bounded and closed, hence compact. It contains all points $\vartheta_n$, since $Z(\vartheta_{n+1}, y) \leq Z(\vartheta_1, y) < \infty$. Hence the sequence $\vartheta_n$; $n = 1, 2, \dots$ has limit points, and the sequence $Z(\vartheta_n, y)$; $n = 1, 2, \dots$ which is non-increasing and bounded from below, has a limit as well.

Let us denote by $\tilde{\vartheta}$ one limit point, and let $\vartheta_{n_k}$; $k = 1, 2, \dots$ be the subsequence converging to it. Suppose that $v(\vartheta) \neq 0$. We have

$$v^T(\tilde{\vartheta})a(\tilde{\vartheta}) = -v^T(\tilde{\vartheta})M(\tilde{\vartheta})v(\tilde{\vartheta}) \, : = d < 0 .$$

Hence for $k > k_0$ we have

$$v^T(\vartheta_{n_k})a(\vartheta_{n_k}) < d/2 .$$

Using the Taylor formula, we obtain for every $\lambda > 0$.

(3.4)

$$Z(\vartheta_{n_k} + \lambda v(\vartheta_{n_k}), y) - Z(\vartheta_{n_k}, y) = 2\lambda a(\vartheta_{n_k}) v(\vartheta_{n_k}) + (\lambda^2/2) v^T(\vartheta_{n_k}) \nabla_\vartheta \nabla_\vartheta^T Z(\vartheta^{\#}, y) v(\vartheta_{n_k})$$

for some $\vartheta^{\#}$. The second term of the right-hand side is bounded, the first is bounded above by $d$, hence for some small $\lambda_0 > 0$ the right-hand side of (3.4) is bounded above by a number $e < 0$. Thus

$$Z(\vartheta_{n_k+1}, y) - Z(\vartheta_{n_k}, y) < Z(\vartheta_{n_k} + \lambda_0 v(\vartheta_{n_k}), y) - Z(\vartheta_{n_k}, y) \leq e$$

which contradicts to the fact that the non-increasing sequence $Z(\vartheta_n, y)$; $n = 1, 2, \dots$ is bounded from below. Hence the assumption $v(\tilde{\vartheta}) \neq 0$ is wrong. Consequently, for every limit point $\tilde{\vartheta}$ we have $v(\tilde{\vartheta}) = 0$, $a(\tilde{\vartheta}) = 0$, and the statements i), ii) and iii) hold. Further

$$\|\vartheta_{n+1} - \vartheta_n\| \leq \|v(\vartheta_n)\| \to 0\,,$$

hence iv) is proven. □

## 4. The Probability Density of the Estimator

Closely related to the presented geometrical analysis is the derivation of the probability density of $\hat{\vartheta}$. For a particular case (case a) in Section 1) a similar approach has been presented in [**7**].

In the whole section we suppose that $M(\vartheta) + G(\vartheta)$ is a positive definite matrix.

For every $\vartheta \in \operatorname{supp}\pi$ take $N - m$ vectors $w_i(\vartheta) \in R^N$ such that for every $i, j, k$ we have

$$\langle w_i(\vartheta), \partial\eta(\vartheta)/\partial\vartheta_k \rangle_\Sigma = 0\,,$$
$$\langle w_i(\vartheta), w_j(\vartheta) \rangle_\Sigma = \delta_{ij}$$

(an orthonormal basis of the ancillary space $\mathscr{B}(\vartheta)$). Denote by $C_i(\vartheta)$ the matrices (the components of the second fundamental form of the surface $\mathscr{E}$)

$$C_i(\vartheta) := w_i(\vartheta) \Sigma^{-1} H(\vartheta)\,.$$

Let us denote

$$S(\vartheta) := \{a \in R^{N-m} : \sup_{v \in R^m, \|v\| = 1} \sum_i a_i [v^T C_i(\vartheta) v] / v^T (M(\vartheta) + G(\vartheta)) v \leq 1\}\,.$$

For every $y \in B(\vartheta)$ denote

$$a_i(y) := \langle y - \eta(\vartheta), w_i(\vartheta) \rangle_\Sigma\,,$$
$$a(y) := (a_1(y), \dots, a_{N-m}(y))^T\,.$$

**Lemma 5.** *If $y \in \mathscr{B}(\vartheta)$, $a(y) \in S(\vartheta)$, then the matrix $\nabla_\vartheta \nabla_\vartheta^T Z(\vartheta, y)$ is positive definite.*

*Proof.* From (2.5) we obtain (in the abbreviated notation)

$$v^T(\nabla\nabla^T Z(y))v = -\sum_i a_i(y)(v^T C_i v) + v^T(M+G)v\,.$$

Hence $a(y) \in S(\vartheta) \Rightarrow v^T(\nabla\nabla^T Z(y))v > 0$ for every $v \neq 0$. □

**Note.** The set $\{y \in \mathscr{B}(\vartheta)\colon \|(I-P)(y-\eta)\|_\Sigma < [K_\pi(\vartheta)]^{-1}\}$ considered in Lemma 2 is a subset of the set $\{y \in \mathscr{B}(\vartheta)\colon a(y) \in S(\vartheta)\}$ considered in Lemma 5.

Let $\overline{\vartheta}$ be the true value of $\vartheta$. Denote by $\psi(\vartheta)$ the auxiliary vector

$$\psi(\vartheta) \,:=\, \eta(\overline{\vartheta}) + P(\vartheta)[\eta(\vartheta) - \eta(\overline{\vartheta})] - u(\vartheta)\,. \tag{4.1}$$

From (2.2) we obtain that for every $y \in \mathscr{B}_\vartheta$ we can write

$$\psi(\vartheta) - y = [I - P(\vartheta)][\eta(\overline{\vartheta}) - y]\,. \tag{4.2}$$

Geometrically, $\psi(\vartheta)$ is the orthogonal projection of the point $\eta(\overline{\vartheta})$ onto the plane $\mathscr{B}_\vartheta$. From (4.2) it follows that we can write

$$y = \psi(\hat{\vartheta}) + \sum_{i=1}^{N-m} b_i w_i(\hat{\vartheta})\,, \tag{4.3}$$

with

$$\begin{aligned} \hat{\vartheta} &= \hat{\vartheta}(y)\,, \\ b_i \,:=\, b_i(y) \,&:=\, \Big\langle y - \psi(\hat{\vartheta}), w_i(\hat{\vartheta})\Big\rangle_\Sigma\,. \end{aligned} \tag{4.4}$$

Evidently, $a_i(y) = b_i(y) + \langle \psi(\vartheta) - \eta(\vartheta), w_i(\vartheta)\rangle_\Sigma$ for every $y \in \mathscr{B}(\vartheta)$. Denote $S^*(\vartheta) \,:=\, \{b(y)\colon\, a(y) \in S(\vartheta)\}$. In the proof of Theorem 3 we prefer $b(y)$ to $a(y)$ because of the important equality (4.10).

**Theorem 3.** *Let $p_\pi(\hat{\vartheta}|\overline{\vartheta})$ be the exact probability density of the random variable $\hat{\vartheta} = \hat{\vartheta}(y)$. Then for every $\vartheta \in \operatorname{supp}\pi$ we have*

$$p_\pi(\hat{\vartheta}|\overline{\vartheta}) = q_\pi(\hat{\vartheta}|\overline{\vartheta}) E_{\hat{\vartheta}}[\det\{I + D(\hat{\vartheta}, b)[Q(\hat{\vartheta}, \overline{\vartheta}) + G(\hat{\vartheta})]^{-1}\}]\,. \tag{4.5}$$

*where*

$$q_\pi(\hat{\vartheta}|\overline{\vartheta}) \; := \; \frac{\det[Q(\hat{\vartheta},\overline{\vartheta}) + G(\hat{\vartheta})]}{(2\pi)^{m/2}\det^{1/2} M(\hat{\vartheta})} \exp\left\{-\frac{1}{2}\|P(\hat{\vartheta})[\eta(\hat{\vartheta}) - \eta(\overline{\vartheta}) - u(\hat{\vartheta})]\|_\Sigma^2\right\} \tag{4.6}$$

$$Q(\hat{\vartheta},\overline{\vartheta}) \; := \; M(\hat{\vartheta}) + [\eta(\hat{\vartheta}) - \eta(\overline{\vartheta})]^T[I - P(\hat{\vartheta})]H(\hat{\vartheta}) \,,$$

$$\{D(\vartheta,b)\}_{ij} \; := \; -\sum_{k=1}^{N-m} b_k \left\langle w_k(\vartheta), \partial^2\eta(\vartheta)/\partial\vartheta_i\partial\vartheta_j\right\rangle_\Sigma = -\sum_{k=1}^{N-m} b_k\{C_k(\vartheta)\}_{ij} \,,$$

$$E_{\hat{\vartheta}}(\cdot) \; := \; \int_{S^*(\hat{\vartheta})} (.)(2\pi)^{-(N-m)/2}\exp\{-\|b\|_I^2/2\}\, d\mu(b) \,,$$

*and where $\mu$ is the Lebesgue measure in $R^{N-m}$.*

*Proof.* Denote by $g(\hat{\vartheta},b)$ the right-hand side of (4.3). From (4.3), (4.4) it follows that the mapping

$$g : (\hat{\vartheta}, b) \in \bigcup_{\vartheta\in\operatorname{supp}\pi} \{\vartheta\}\times S(\vartheta) \to g(\hat{\vartheta},b) \in R^N$$

is a bijection (up to a set of Lebesgue measure zero in $R^N$). It is also differentiable, and its Jacobian is equal to (cf. [**6**, eq. (19)])

$$|\det[\nabla g(\hat{\vartheta},b)]| = \frac{\det[D(\hat{\vartheta},b) + \nabla_\vartheta\psi^T(\hat{\vartheta})\Sigma^{-1}J(\hat{\vartheta})]}{\det^{1/2} M(\hat{\vartheta})}\det{}^{1/2}(\Sigma) \,. \tag{4.7}$$

The joint probability density of $\hat{\vartheta}$ and $b$, $p(\hat{\vartheta},b|\overline{\vartheta})$ is equal to

$$p(\hat{\vartheta},b|\overline{\vartheta}) = |\det[\nabla g(\hat{\vartheta},b)]|(2\pi)^{N/2}\det{}^{-1/2}(\Sigma)\exp\{-(1/2)\|y - \eta(\overline{\vartheta})\|_\Sigma^2\}_{y=g(\hat{\vartheta},b)} \,, \tag{4.8}$$

and the required density of $\hat{\vartheta}$ is equal to

$$p_\pi(\hat{\vartheta}|\overline{\vartheta}) = \int_{S^*(\hat{\vartheta})} p(\hat{\vartheta},b|\overline{\vartheta})\, d\mu(b) \,. \tag{4.9}$$

From (4.2) it follows by the theorem of Pythagoras that

$$\begin{aligned} \|y - \eta(\overline{\vartheta})\|_\Sigma^2 &= \|y - \psi(\hat{\vartheta})\|_\Sigma^2 + \|\psi(\hat{\vartheta}) - \eta(\overline{\vartheta})\|_\Sigma^2 \\ &= \|b\|_I^2 + \|P(\hat{\vartheta})[\eta(\hat{\vartheta}) - \eta(\overline{\vartheta})] - u(\hat{\vartheta})\|_\Sigma^2 \,. \end{aligned} \tag{4.10}$$

Let us multiply (4.1) from the left by $[\partial\eta^T(\vartheta)/\partial\vartheta_i]\Sigma^{-1}$. We obtain

$$[\partial\eta^T(\vartheta)/\partial\vartheta_i]\Sigma^{-1}[\psi(\vartheta)-\eta(\vartheta)]+\partial l(\vartheta)/\partial\vartheta_i=0\,.$$

We take the derivative of this with respect to $\vartheta_j$, and we use (4.1) again, to obtain

$$\frac{\partial\eta^T(\vartheta)}{\partial\vartheta_i}\Sigma^{-1}\frac{\partial\psi(\vartheta)}{\partial\vartheta_j}=Q_{ij}(\hat\vartheta,\overline{\vartheta})+G_{ij}(\hat\vartheta)\,.$$

We put this into (4.7), and from (4.8)–(4.10) we obtain the required equality. □

**Remark.** The expression given in eq. (4.6) can be considered as an approximation of the probability density of $\hat\vartheta$ which is easy to compute. It si a direct generalization of the density of the least-squares estimator discussed in [**6**], to the case of a given prior. For a particular purpose it has been applied in [**7**].

## References


1. Amari S., *Differential Geometrical Methods in Statistics*, Lecture Notes in Statistics, No. 28, Springer Verlag, 1985.
2. Bates D. M. and Watts D. G., *Relative curvature measures of nonlinearity (with discussion)*, J. Roy. Statist. Soc. **B 42** (1980), 1–25.
3. Ford I. Kitsos, Ch. P. and Titterington D. M., *Recent advances in nonlinear experimental design*, Technometrics **31** (1989), 49–60.
4. Johansen S., *Functional relations, random coefficients and nonlinear regression with application to kinetic data.*, Lecture Notes in Statistics, No. 22, Springer Verlag, 1984.
5. Kass R. E., *The geometry of asymptotic inference. (with discussion)*, Statistical Science **4** (1989), 188–234.
6. Pázman A., *On formulas for the distribution of the nonlinear L. S. estimates*, Statistics **18** (1987), 3–15.
7. Pázman A. and Pronzato L., *Nonlinear experimental design based on the distribution of estimates*, Jour. Statist. Planning & Inference (1992), (to appear).


A. Pázman, Department of Probability and Statistics, Faculty of Mathematics and Physics, Comenius University, 842 15 Bratislava, Czechoslovakia