Math. Appl. **8** (2019), 101–114
DOI: 10.13164/ma.2019.07

ON A GLOBAL MEASURE OF NONLINEARITY AND ITS APPLICATION IN PARAMETER ESTIMATION IN NONLINEAR REGRESSION

LEONID KHINKIS

Abstract. The theoretical and computational challenges in least squares estimation of parameters in nonlinear regression models are well documented in statistical literature. The measures of nonlinearity are intended to quantify the degree of nonlinearity and to explain the relationship between nonlinearity and statistical properties of a model. A new measure of nonlinearity reflecting the model's global behavior is introduced and discussed in this paper. Two new criteria for global minimum of the sum of squares in nonlinear regression incorporating this measure are presented and illustrated on several published examples.

1. Introduction

It has been widely recognized that nonlinearity may adversely affect different aspects of statistical inference in nonlinear regression. Beale [1] and Bates and Watts [2] were the first to apply some fundamental notions and powerful methods of differential geometry to the study of nonlinear regression models. Specifically, Beale applied the notion of curvature to improve approximate confidence regions while Bates and Watts identified intrinsic and parametric curvatures and developed inferential procedures based on these measures. This work was further advanced in a number of publications including [6] and [7]. One needs to note that these curvature measures are local since they are computed from the partial derivatives of the model evaluated at a specified parameter vector. Thus, these measures reflect local behavior of the model in a neighborhood of the specified parameter and may lack an ability to capture the model's global behavior. This motivates development of methods based on techniques other than the Taylor's expansion in a neighborhood of the specified parameter. Construction of global criteria for minimization of the residual sum of squares (SS) is an important problem in nonlinear least squares estimation. Unlike in linear models, SS may possess multiple minima in nonlinear regression models [4, 8], which raises a question whether the final iteration of any SS minimization routine results in a true least-squares estimate. Chavent [3] and Demidenko [4,5] obtained criteria (sufficient conditions) ensuring that a local minimizer of the residual sum of squares is also a global minimizer. In the absence of a general global function minimization algorithm, sufficient conditions, like in the

MSC (2010): primary 62J02, 62K05, 62P10.
Keywords: nonlinear regression, global minimization, curvature, global measure of nonlinearity.

work cited above, may help to identify a true parameter estimate. Noteworthy, it was shown in [5] that the level of the local unimodality of the sum of squares used in Demidenko's criterion equals to the minimum squared radius of the intrinsic curvature of the nonlinear regression model. Demidenko's contributions [4,5] illustrate the utility of the local curvature measures in development of global methods. Pronzato and Pázman [14] introduced a global measure of nonlinearity and called it an extended measure of intrinsic nonlinearity due to its relationship to the intrinsic curvature of Bates and Watts [2]. This paper introduces a new, geometrically appealing, measure of nonlinearity (MoN) and places it into the context of the existing work. This measure advances methodology relying on the notion of equidistant function originally introduced in [9] and further developed in [10,11]. The needed definitions and terminology are presented briefly in this section. Two new global criteria for minimization of SS based on the mentioned MoN are introduced in Section 2. These criteria are compared to the criteria of Demidenko and Chavent using several examples from the published literature. This is done in Section 3. It is argued in Section 4 that the new MoN is a natural global extension of the intrinsic curvature of Bates and Watts and is also related to the global MoN of Pronzato and Pázman.

Let us consider a nonlinear regression model given by

$$y = \eta(\theta) + \varepsilon, \quad \theta \in \Theta, \quad \mathbb{E}(\varepsilon) = 0, \quad \operatorname{var}(\varepsilon) = \sigma^2 W,$$

where $\theta = (\theta_1, \dots, \theta_m)^t$ is a transposed column vector of unknown parameters. Assume that $\theta \in \Theta$, and that the (known) parameter space Θ is a subset of $\mathbb{R}^m$. Furthermore, $y \in \mathbb{R}^N$ is the vector of observed data, $\varepsilon \in \mathbb{R}^N$ is the error vector, σ is the parameter of the variance component which may, but needs not be known, W is a known positive semi-definite matrix. The parameter σ equals to the standard deviation of the error of an individual observation under the assumption of a constant variance, in which case $W = I$, the identity matrix.

We assume that the errors are normally distributed. The mapping $\eta : \Theta \to \mathbb{R}^N$ is a known, twice continuously differentiable mapping on $\operatorname{int}\Theta$.

The expectation surface is defined as $E_\eta = \{\eta(\theta) : \theta \in \Theta\}$. The least squares estimate (LSE) of θ is

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

and the definition of the squared norm $\|a\|_W^2 = a^t W^{-1} a$ is used. This norm corresponds to the inner product $\langle a, b\rangle_W = a^t W^{-1} b$. In what follows, $\|\cdot\|$ is understood to be $\|\cdot\|_{W=I}$.

Any LSE $\hat{\theta}(y) \in \operatorname{int}\Theta$ satisfies the system of m normal equations (the stationary conditions)

$$\frac{\partial}{\partial \theta}\|y - \eta(\theta)\|_W^2 = 0. \tag{1.1}$$

The linear span of the vectors $\frac{\partial \eta(\theta)}{\partial \theta_i}$, $(i = 1, \dots, m)$, forms a plane, $T(\theta)$, known as the tangent plane to the expectation surface E_η at $\theta \in \Theta$. Define the normal plane $NO(\theta)$ as the hyperplane orthogonal to the expectation surface at the point

$\eta(\theta)$,

$$NO(\theta) := \left\{ n \in \mathbb{R}^N : \left\langle n, \frac{\partial \eta(\theta)}{\partial \theta_i} \right\rangle_W = 0, \, (i = 1, \ldots, m) \right\}.$$

Let $P(\theta)$ be an orthogonal projector onto the tangent plane $T(\theta)$. Then any vector $h \in \mathbb{R}^N$ can be represented as a sum of its mutually orthogonal components,

$$h = P(\theta)h + (I - P(\theta))h,$$

so that $P(\theta)h \in T(\theta)$ while $(I - P(\theta))h \in NO(\theta)$. Let $NO_1(\theta)$ be the set of all unit vectors in $NO(\theta)$, $NO_1(\theta) := \{n \in NO(\theta) : \|n\|_W = 1\}$ and $\mathbb{R}_+$ be the set of all positive real numbers.

A *directional* equidistant function, $t(\theta, \theta_1, n)$ was defined in [9, Eq. (3)] as

$$t(\theta, \theta_1, n) = \frac{\|\eta(\theta_1) - \eta(\theta)\|_W^2}{2\langle n, \eta(\theta_1) - \eta(\theta)\rangle_W}.$$

Here $\theta \in \operatorname{int} \Theta$, $\theta_1 \in \Theta$ and $n \in NO_1(\theta)$. As pointed out in [9], $y = \eta(\theta) + tn$ is equidistant from two different points, $\eta(\theta)$ and $\eta(\theta_1)$ if and only if $\langle n, \eta(\theta_1) - \eta(\theta)\rangle_W > 0$ and $t = t(\theta, \theta_1, n)$. Here $t \in \mathbb{R}_+$, $\theta \in int\, \Theta$, $\theta_1 \in \Theta$ and $n \in NO_1(\theta)$.

The equidistance property refers to the equality

$$\|y - \eta(\theta)\|_W = \|y - \eta(\theta_1)\|_W.$$

This property will not hold for any $y \in \mathbb{R}^N$, $\theta \in int\, \Theta$, $\theta_1 \in \Theta$, $n \in NO_1(\theta)$ and $t \in \mathbb{R}_+$ such that $y = \eta(\theta) + tn$ and $\langle n, \eta(\theta_1) - \eta(\theta)\rangle_W \leq 0$. By Theorem 1 from [9], for a fixed $\theta_1 \in int\, \Theta$, $t(\theta, \theta_1, n)$ is the supremum of the values $d \in \mathbb{R}_+$ such that $y = \eta(\theta) + dn$ results in the unique least square estimate $\hat{\theta}(y) = \theta$. Let's define $N_1(\theta, \theta_1)$ as

$$N_1(\theta, \theta_1) = \{n \in NO_1(\theta) : \langle n, \eta(\theta_1) - \eta(\theta)\rangle_W > 0\}.$$

The equidistant function $t(\theta, \theta_1)$ is defined in [11] as

$$t(\theta, \theta_1) = \inf_{n \in N_1(\theta, \theta_1)} t(\theta, \theta_1, n). \tag{1.2}$$

By definition, $t(\theta, \theta_1) = +\infty$ if $N_1(\theta, \theta_1) = \emptyset$.

The Proposition 2.1 in [11] establishes that

$$t(\theta, \theta_1) = \frac{\|\eta(\theta_1) - \eta(\theta)\|_W^2}{2\|(I - P(\theta))(\eta(\theta_1) - \eta(\theta))\|_W}, \quad \theta \in \operatorname{int} \Theta, \, \theta_1 \in \Theta. \tag{1.3}$$

Define

$$d(\theta) = \inf_{\theta_1 \in \Theta} t(\theta, \theta_1), \quad \theta \in \operatorname{int} \Theta. \tag{1.4}$$

The intrinsic curvature of a nonlinear regression model is defined in [13] as

$$C_{int}(\theta) = \sup_{v \in \mathbb{R}^m \setminus \{0\}} C_{int}(\theta, v) \tag{1.5}$$

where the directional curvature is

$$C_{int}(\theta, v) = \frac{\|(I - P(\theta))v^t H(\theta) v\|}{v^t M(\theta) v}, \tag{1.6}$$

and

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

is an orthogonal projector onto the tangent plane $T(\theta)$ introduced above, while the expressions below represent the information matrix, Jacobian, and Hessian, respectively:

$$M(\theta) = J^t(\theta)W^{-1}J(\theta),$$
$$J(\theta) = \frac{\partial\eta(\theta)}{\partial\theta^t}, \quad H(\theta) = \frac{\partial^2\eta(\theta)}{\partial\theta\partial\theta^t}.$$

Geometrically, $C_{int}(\theta)$ represents the maximal curvature of the geodesic curves on E passing through $\theta \in int\,\Theta$. The radius of intrinsic curvature,

$$R_{int}(\theta) := \frac{1}{C_{int}(\theta)},$$

represents the infimum of $t(\theta, \theta_1)$ when θ_1 approaches θ along all possible directions.

2. Criteria for global minimum of the sum of squares

As mentioned in the Introduction, $\|y - \eta(\theta)\|^2_W$ may have multiple local minimizers. Since most algorithms for LS estimation only perform a local search, it is then difficult to certify that the minimizer obtained is the global one. Two existing criteria serving this purpose are presented below along with two new ones.

Criterion 1: The following is a reformulation of Theorem 7.3 from [3].

Assume that $\eta(\cdot)$ is twice differentiable in $\operatorname{int}\Theta$, where Θ is a compact subset of $\mathbb{R}^m$, and that the Jacobian $J(\theta)$ has full rank m for any $\theta \in \operatorname{int}\Theta$. Define

$$\alpha_\eta = \left(\min_{\theta\in\Theta} \lambda_{\min}[M(\theta)]\right)^{1/2},$$

$$\beta_\eta = \max_{\theta\in\Theta} \max_{u\in\mathbb{R}^m, \|u\|=1} \left(\sum_{i=1}^N \left[u^t \frac{\partial^2\eta(x_i,\theta)}{\partial\theta\partial\theta^t} u\right]^2\right)^{1/2},$$

and $\operatorname{diam}(\Theta) = \max_{\theta,\theta'\in\Theta} \|\theta' - \theta\|$. If

$$\operatorname{diam}(\Theta) < 2\sqrt{2}\frac{\alpha_\eta}{\beta_\eta},$$

then for any y such that the distance

$$d(y, E_\eta) = \min_{\theta\in\Theta} \|y - \eta(\theta)\| < \frac{\alpha_\eta^2}{\beta_\eta} - \frac{\beta_\eta}{8}[\operatorname{diam}(\Theta)]^2, \tag{2.1}$$

the LS estimator has a unique global minimizer $\hat{\theta}(y)$ depending continuously on y, and there exists no other local minimizers.

Criterion 2: ([5, Thm. 2]) The global criterion is formulated in terms of the level set defined as

$$L(S_*) = \{\theta \in \Theta : S(\theta) < S_*\}.$$

The upper local unimodality level for the sum of squares of a nonlinear regression model is defined in [5] as $\bar{S}_{LU} = \min_{\theta\in\Theta} R^2_{int}(\theta)$. A local unimodality level, S_{LU}, is defined as any number less or equal to $\bar{S}_{LU}$. Demidenko's Theorem 2 in [5] states that there is at most one local minimum in each connected component of the level set $L(\bar{S}_{LU})$. Moreover, if $\hat{\theta}$ is a local minimizer such that $S(\hat{\theta}) < \bar{S}_{LU}$ and the level set $L(\bar{S}_{LU})$ is connected, then $\hat{\theta}$ is the global minimizer.

Criterion 3: Let $\theta^* \in \operatorname{int}\Theta$ satisfies the stationary conditions (1.1) together with $\sqrt{S(\theta^*)} < d(\theta^*)$. Then θ^* is a global minimizer.

Proof. Assume that θ^* is not a global minimizer. Then there exists $\theta_1 \in \Theta$ such that $S(\theta_1) < S(\theta^*)$. Let $n = \dfrac{y - \eta(\theta^*)}{\|y - \eta(\theta^*)\|_W}$. The function

$$k(l) = \|\eta(\theta^*) + ln - \eta(\theta_1)\|_W - l$$

is a continuous function of l on $\left[0, \sqrt{S(\theta^*)}\right]$. Moreover, since $k(0) = \|\eta(\theta^*) - \eta(\theta_1)\|_W > 0$, and

$$k(\sqrt{S(\theta^*)}) = \|\eta(\theta^*) + \sqrt{S(\theta^*)} \cdot n - \eta(\theta_1)\|_W - \sqrt{S(\theta^*)} = \sqrt{S(\theta_1)} - \sqrt{S(\theta^*)} < 0,$$

there exists $0 < t < \sqrt{S(\theta^*)}$ such that $k(t) = 0$. Then, $z = \eta(\theta^*) + tn$ is equidistant from two distinct points, $\eta(\theta^*)$ and $\eta(\theta_1)$ and, as explained in Section 1, $n \in N_1(\theta^*, \theta_1)$ and $t = t(\theta^*, \theta_1, n)$. It follows from Eqs. (1.4) and (1.2) that

$$d(\theta^*) \leq t(\theta^*, \theta_1) \leq t(\theta^*, \theta_1, n) < \sqrt{S(\theta^*)}.$$

This inequality contradicts the assumption that $\sqrt{S(\theta^*)} < d(\theta^*)$. □

Criterion 4: Let $\theta^* \in \operatorname{int}\Theta$ satisfies the stationary conditions (1.1) and

$$\sqrt{S(\theta^*)} < d, \quad \text{where} \quad d = \inf_{\theta \in \operatorname{int}\Theta} d(\theta) = \inf_{\theta \in \operatorname{int}\Theta, \theta_1 \in \Theta} t(\theta, \theta_1).$$

Then θ^* is a global minimizer.

Proof. This is a rather trivial corollary from Criterion 3 since, under assumptions of Criterion 4, $\sqrt{S(\theta^*)} < d \leq d(\theta^*)$. □

Clearly, Criterion 4 is weaker than Criterion 3. However, the upper bound on $\sqrt{S(\theta^*)}$ being not dependent on θ^*, could be a desirable feature.

3. Comparison of different criteria

Pronzato and Pázman [15] offer two examples highlighting some issues concerning practicality of both Criteria 1 and 2. Since, as shown in [5], Criterion 2 is an improvement over Demidenko's other criterion [4] and [15, Thm. 7.4], we refer to Criterion 2 in our analysis.

For simplicity, in the following examples we use $W = I$, but the results apply in general case.

Example 3.1. ([15, Exm. 7.5]) Consider the following one-parameter model:

$$\eta(x, \theta) = \theta\{x\}_1 + \theta^2\{x\}_2$$

with two observations at the design points $x_1 = (1, 0)$, $x_2 = (0, 1)$. Assume that $0 \in \Theta \subset \mathbb{R}$. Thus, $\eta(\theta) = (\theta, \theta^2)^t$. Direct calculations result in $M(\theta) = 1 + 4\theta^2$, $\alpha_\eta = 1$, $\beta_\eta = 2$ and

$$R_{int}(\theta) = \frac{(1 + 4\theta^2)^{3/2}}{2}. \tag{3.1}$$

Clearly, $R^2_{int}(\theta)$ reaches its minimum at 0, meaning that $\bar{S}_{LU}$ defined in Criterion 2 is 1/4.

The expectation surface $E_\eta = \{(\theta, \theta^2)^t : \theta \in \mathbb{R}\}$ is a parabola shown in Figure 1. The bound on $d(y, E_\eta)$ given by (2.1) equals $b(\theta) = \frac{1}{2} - \frac{(\mathrm{diam}(\Theta))^2}{4}$, which defines a tube around E_η (see the blue dashed-line curves in Figure 1 obtained for $\mathrm{diam}(\Theta) = 1$).

Criterion 2 defines a tube around E_η with the boundary $b = \frac{1}{2}$ (matching $\mathrm{diam}(\Theta) = 0$ in Criterion 1) shown as red dotted-line curves in Figure 1, while criterion 1 defines a smaller tube of a decreasing size as $\mathrm{diam}(\Theta)$ increases.

We will now compute the equidistant function $t(\theta, \theta_1)$ given by Eq. (1.2) and $d(\theta^\star)$ given by Eq. (1.4). The tangent vector $t(\theta)$ to the expectation surface E_η is given by $(1, 2\theta)$; the unit normal vector $n \in NO_1(\theta)$ is given by $n = \frac{(-2\theta,1)}{\sqrt{1+4\theta^2}}$. Then

$$\begin{aligned} t(\theta, \theta_1, n) &= \frac{\|\eta(\theta_1) - \eta(\theta)\|^2}{2\langle n, \eta(\theta_1) - \eta(\theta)\rangle} = \frac{((\theta_1^2 - \theta^2)^2 + (\theta_1 - \theta)^2)\sqrt{1+4\theta^2}}{2(-2\theta(\theta_1 - \theta) + \theta_1^2 - \theta^2)} \\ &= \frac{(\theta_1 - \theta)^2((\theta_1 + \theta)^2 + 1)\sqrt{1+4\theta^2}}{2(\theta_1 - \theta)^2} = \frac{((\theta_1 + \theta)^2 + 1)\sqrt{1+4\theta^2}}{2}. \end{aligned}$$

Since $NO_1(\theta)$ is one-dimensional, then

$$t(\theta, \theta_1) = t(\theta, \theta_1, n) = \frac{((\theta_1 + \theta)^2 + 1)\sqrt{1+4\theta^2}}{2}. \tag{3.2}$$

Then $d(\theta) = \inf_{\theta_1 \in \Theta} t(\theta, \theta_1) = t(\theta, -\theta) = \frac{\sqrt{1+4\theta^2}}{2}$ while $d = \inf_{\theta \in \mathrm{int}\,\Theta} d(\theta) = d(0) = \frac{1}{2}$.

The quantity $d(\theta)$ is the distance between a point $A(\theta, \theta^2)$ on the expectation surface and the point $M(0, \theta^2 + \frac{1}{2})$ where the normal line AM meets the vertical axis $y_1 = 0$. Criterion 3 specifies a region above parabola $y = \frac{\theta^2}{4} - \frac{1}{2}$ (given by the green dash dot line-curve in Figure 1) which contains the tube bounded by the red dotted line curves (Figure 1) specified by Criterion 2. This means that Criterion 3 is stronger than Criterion 2.

Criterion 4 uses the same quantity $d = \sqrt{\bar{S}_{LU}} = \frac{1}{2}$ as Criterion 2. Yet it is easier to implement since Criterion 2 additionally requires connectivity of the level set $L(\bar{S}_{LU})$ that needs to be verified.

Example 3.2. ([15, Exm. 7.6]) The previous example is modified by changing $\eta(x, \theta)$ for negative θ,

$$\eta(x, \theta) = (\theta\{x\}_1 + \theta^2\{x\}_2)\mathbb{1}_{\mathbb{R}^+}(\theta) + (\sin\theta\{x\}_1 + 2(1 - \cos\theta)\{x\}_2)\mathbb{1}_{\mathbb{R}^-}(\theta)$$

with $\Theta = [\gamma, \theta_{\max}]$ or $\Theta = [\gamma, +\infty), \gamma > -2\pi$. Since Criterion 1 requires compactness of Θ, we will use $\Theta = [\gamma, \theta_{\max}]$ when discussing this criterion.

The design points $x_1 = (1, 0)$ and $x_2 = (0, 1)$ remain as in the previous example. One can verify that the vector function $\eta(\theta)$ presented in Figure 2 is twice continuously differentiable in the interior of Θ.

Following [15], $M(\theta) = \cos^2\theta + 4\sin^2\theta$ for $\theta \leq 0$; $M(\theta) = 1 + 4\theta^2$ for $\theta > 0$, $\alpha_\eta = 1$, $\beta_\eta = 2$ as in Example 3.1. Criterion 1 imposes a strict restriction on

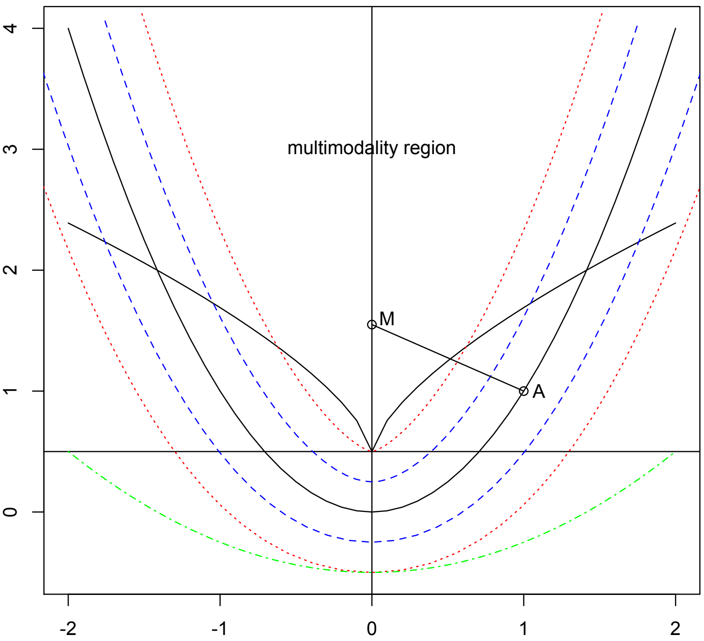

Figure 1. Regions in the sample space specified by different criteria in Example 3.1.

diam(Θ) due to inequality (2.1). In order for (2.1) be meaningful, its right-hand side must be positive:

$$\frac{\alpha_\eta^2}{\beta_\eta} - \frac{\beta_\eta}{8}[\mathrm{diam}(\Theta)]^2 > 0,$$

which, given the specified values of $\alpha_\eta = 1$ and $\beta_\eta = 2$, leads to $\mathrm{diam}(\Theta) < \sqrt{2}$ or, equivalently, $\theta_{\max} < \gamma + \sqrt{2}$. Thus, Criterion 1 is not applicable if $\theta_{\max} \geq \gamma + \sqrt{2}$.

The values of $R_{int}(\theta)$ are computed as

$$R_{int}(\theta) = \frac{(4 - 3\cos^2\theta)^{3/2}}{2} \quad \text{when} \quad \theta \leq 0.$$

When $\theta > 0$, $R_{int}(\theta)$ remains the same as in (3.1). So, $\min_{\theta\in\Theta} R_{int}(\theta) = \frac{1}{2}$, same as in Example 3.1 resulting in $\bar{S}_{LU} = \frac{1}{4}$. As explained in [15, Exm. 7.6], although the expectation surface E_η almost overlaps, performance of Criterion 2 is similar to that exhibited in Example 3.1. The reader is referred to [15] for details.

Although the expectation surface E_η folds over itself, Criterion 2 is not responsive to this behavior. Consequently, it leads to a false statement based on the value of $\bar{S}_{LU}$ alone. Connectedness of $L(\bar{S}_{LU})$ needs to be verified which, in general, makes Criterion 2 less practical. As pointed out in [15] and illustrated by this example, "the information provided by the (local) curvature of the model is clearly not enough to measure the difficulty of the estimation of its parameters caused by an expectation surface that folds over itself".

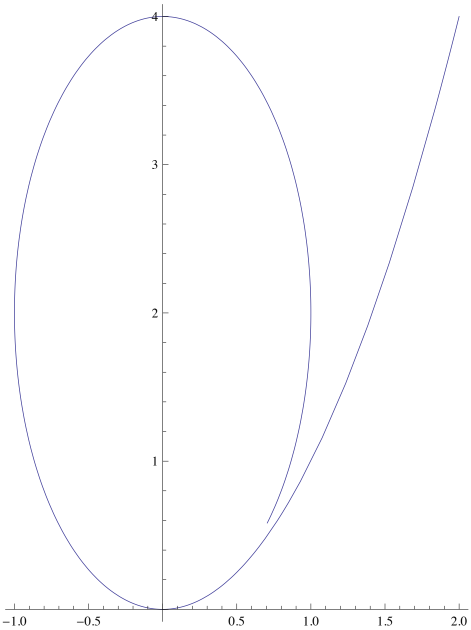

Figure 2. Expectation surface in Example 3.2.

Assume $\theta \geq 0$ and $\theta_1 > 0$. Then, according to (3.2),

$$\inf_{\theta_1 > 0} t(\theta, \theta_1) = t(\theta, 0) = \frac{(1+\theta^2)\sqrt{1+4\theta^2}}{2}.$$

When $\theta_1 < 0$, (1.2) results in

$$\begin{aligned} t(\theta, \theta_1) &= t(\theta, \theta_1, n) = \frac{\|\eta(\theta_1) - \eta(\theta)\|^2}{2\langle n, \eta(\theta_1) - \eta(\theta_{})\rangle} \\ &= \frac{((\sin\theta_1 - \theta)^2 + (2(1-\cos\theta_1) - \theta^2)^2)\sqrt{1+4\theta^2}}{2(\theta^2 - 2\theta\sin\theta_1 + 2 - 2\cos\theta_1)}. \end{aligned}$$

The graph of $t(\theta, \theta_1)$ is presented in Figure 3. The graph of $t(\theta, \theta_1)$, with $\theta = .75$ selected for illustration purpose, is given in Figure 4. The graph has a discontinuity at $\theta_1 = 0$, two local minima at $\theta_1 = -3.989$ and $\theta_1 = -5.43$, and the global minimum of

$$d(0.75) = \min_{\theta_1 \in (-2\pi, \infty)} t(0.75, \theta_1) = t(0.75, -3.989) = 0.0002.$$

However, if $\Theta = [\gamma, \infty)$ with $\gamma > -3.989$ then

$$d(0.75) = t(0.75, \gamma).$$

It is clear that $d = \min_{\theta \in (-2\pi, \infty)} d(\theta) = 0$ meaning that Criterion 4 is not helpful (as warranted by the example), yet Criterion 3 is applicable.

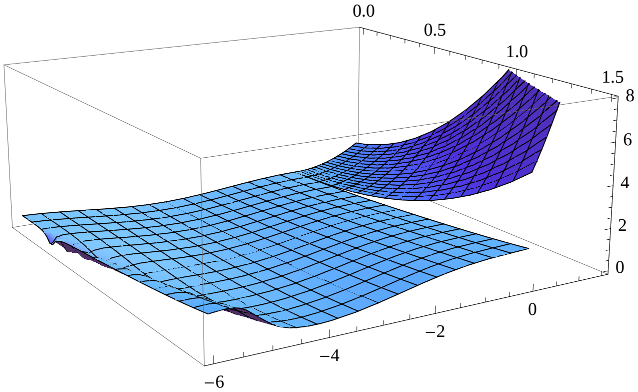

Figure 3. The graph of $t(\theta, \theta_1)$.

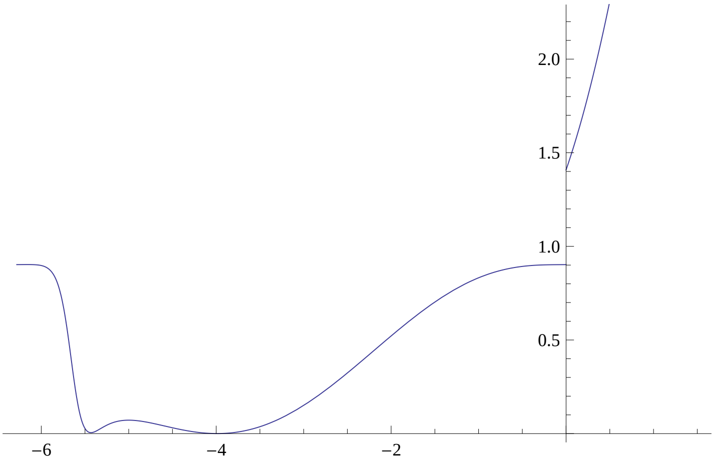

Figure 4. The graph of $t(0.75, \theta_1)$.

Example 3.3. Models with a single linear parameter a.

Calculation of t when $\theta = (a, \gamma)^t$ and $\eta(a, \gamma) = a\varphi(\gamma)$. In this case,

$$\frac{\partial \eta}{\partial a} = \varphi(\gamma), \quad \frac{\partial \eta}{\partial \gamma} = a\varphi'(\gamma).$$

In this example, $P(a, \gamma)k = P(1, \gamma)k$ for $\forall\ k \in \mathbb{R}^n$. In what follows, $P(\gamma) := P(1, \gamma)$. Also, $P(\gamma)\varphi(\gamma) = \varphi(\gamma)$ since $\varphi(\gamma) \in T(\theta)$. So,

$$t(a, \gamma, a_1, \gamma_1) = \frac{a_1^2\|\varphi(\gamma_1)\|^2 - 2a_1 \cdot a\langle\varphi(\gamma_1), \varphi(\gamma)\rangle + a^2\|\varphi(\gamma)\|^2}{2a_1\|(I - P(\gamma))\varphi(\gamma_1)\|}. \tag{3.3}$$

Note that the derivative of a function given by $f(x) = \frac{ex^2+bx+c}{dx}$ is computed as $f'(x) = \frac{ex^2-c}{dx^2}$, meaning that its critical numbers are $x = \pm\sqrt{\frac{c}{e}}$. This translates into

$$a_1 = \sqrt{\frac{a^2\|\varphi(\gamma)\|^2}{\|\varphi(\gamma_1)\|^2}} = a\frac{\|\varphi(\gamma)\|}{\|\varphi(\gamma_1)\|} \tag{3.4}$$

being a critical number for $t(a_1) = t(a, \gamma, a_1, \gamma_1)$ as a function of a_1, with fixed (a, θ, θ_1).

It then follows that $\min_{a_1 \in (0,l)} t(a_1)$ is achieved at $a_1^* = a\frac{\|\varphi(\gamma)\|}{\|\varphi(\gamma_1)\|}$ when $l > a_1^*$ or at $a_1^* = l$ when $l \leq a_1^*$.

Assume that $l > a_1^*$; then, by substituting (3.4) into (3.3), one obtains

$$\min_{a_1 \in (0,l)} t(a_1) = a \cdot g(\gamma, \gamma_1),$$

where

$$g(\gamma, \gamma_1) = \frac{\|\varphi(\gamma)\| \cdot \|\varphi(\gamma_1)\| - \langle \varphi(\gamma_1), \varphi(\gamma)\rangle}{\|(I - P(\gamma))\varphi(\gamma_1)\|}.$$

The Michaelis–Menten model given by

$$\eta(a, \gamma, x) = \frac{a \cdot x}{\gamma + x}$$

represents a specific example belonging to the class of models considered in Example 3.3. The model could be used to describe a physiological response, $\eta(a, \gamma, x)$, as a function of a drug concentration, x, and contains two parameters, the maximum response a and γ, the concentration resulting in 50% of the maximum response.

The model was originally developed by Michaelis and Menten [12] to describe the metabolism of an agent by a reaction rate, in which case the response represents the velocity of an enzyme-substrate reaction.

Here

$$\varphi(\gamma) = \left[\frac{x_i}{x_i + \gamma}\right]^t, \quad i = 1, \ldots, n.$$

This model is discussed in the context of global curvature in [10].

Example 3.4. Consider

$$\eta(\beta, a, \gamma) = X\beta + a\varphi(\gamma)$$

where X is a $n \times k$ matrix, β is a $k \times 1$ vector of linear parameters, γ is a $l \times 1$ vector, a is a scalar, $\varphi(\gamma)$ is a $n \times 1$ vector, and $k + l + 1 = m$. Then $\theta = (\beta, a, \gamma)^t \in \mathbb{R}^m$ is the vector of estimated parameters, and Eq. (1.3) becomes

$$t(\theta, \theta_1) = \frac{\|X\beta_1 + a_1\varphi(\gamma_1) - X\beta - a\varphi(\gamma)\|^2}{2\|(I - P(\theta))(a_1\varphi(\gamma_1) - a\varphi(\gamma))\|}. \tag{3.5}$$

Since $T(\theta)$ is generated by $\{\varphi(\gamma), \partial_\gamma\varphi(\gamma), u_j\}$ where u_j is the j-th column of matrix X, the following equalities hold:

$$P(\theta)X\beta = X\beta,\ P(\theta)\varphi(\gamma) = \varphi(\gamma),\ P(\theta)X\beta_1 = X\beta_1.$$

Moreover, $P(\theta) = P(1, \ldots, 1, \gamma) = P(\gamma)$. Thus, the right hand side of (3.5) becomes

$$\frac{\|a_1\varphi(\gamma_1) - X(\beta - \beta_1) - a\varphi(\gamma)\|^2}{2|a_1|\|(I - P(\gamma))\varphi(\gamma_1)\|}. \tag{3.6}$$

Minimization of Eq. (3.6) over β and β_1 is a linear regression problem, with the minimum realized at

$$\beta - \beta_1 = (X^t X)^{-1} X^t (a_1 \varphi(\gamma_1) - a\varphi(\gamma)),$$

meaning that

$$\min_{\beta,\beta_1 \in \mathbb{R}^k} t(\theta, \theta_1) = \frac{\|a_1(I-H)\varphi(\gamma_1) - a(I-H)\varphi(\gamma)\|^2}{2|a_1|\|(I-P(\gamma))\varphi(\gamma_1)\|}. \tag{3.7}$$

Here $H = X(X^t X)^{-1} X^t$, the well-known hat matrix.

Minimizing Eq. (3.7) over a_1 is a problem solved in Example 3.3 leading to

$$\min_{\beta,\beta_1 \in \mathbb{R}^k, a_1 \in (0,\infty)} t(\theta, \theta_1) = a \cdot g_1(\gamma, \gamma_1),$$

where

$$g_1(\gamma,\gamma_1) = \frac{\|a_1(I-H)\varphi(\gamma_1)\|\|(I-H)\varphi(\gamma)\| - \langle (I-H)\varphi(\gamma_1), (I-H)\varphi(\gamma)\rangle}{\|(I-P(\gamma))\varphi(\gamma_1)\|}.$$

The Log-Gompertz model, a popular model used in biology and medicine, particularly for modeling tumor growth, will be used as an illustration. The estimation of the Gompertz model reduces to a nonlinear regression [5]:

$$y_i = \beta + a e^{\gamma x_i} + \varepsilon_i, \quad i = 1, \ldots, N$$

where y_i is the logarithm of an observation, x_i is the time point of observation. Then $X = (1,\ldots,1)^t$, $X^t X = n$, $H = \frac{1}{n} X X^t = \frac{1}{n} U$, where U is the unit matrix. We assume that both a and γ are negative, which is often the case in applications and that $x_t = t$. It follows that

$$\min_{a,a_1 \in (-\infty,0)} t(\theta,\theta_1) = -a g_1(\gamma,\gamma_1).$$

Let us consider the design points $x_i = 5$, 10, 15, 20, 25, 30 used with parameter values $a = -5$, $\gamma = -0.1$, $\beta = 6$. The values of $g(\gamma_1) = g_1(-0.1, \gamma_1)$ are presented in Table 1.

Table 1. Log-Gompertz model.

γ_1	-.01	-.02	-.05	-.1	-.2	-.3	-.5	-.8	-1	-1.2	-1.5
$g(\gamma_1)$	.13	.216	.32	.3	.18	.104	.038	.0084	.003	.001	.00025

4. An extended measure of intrinsic nonlinearity

As illustrated in Examples 3.1 and 3.2, intrinsic curvature given by Eq. (1.6) may not be able to capture the global behavior of a nonlinear model. This necessitates a concept of intrinsic curvature which does that. Pronzato and Pázman ([15], p. 207) defined $K_{int,\alpha}(\theta) = \sup_{\theta_1 \in int(\Theta)} K_{int}(\theta,\theta_1)$ as such measure, where

$$K_{int,\alpha}(\theta,\theta_1) = 2\frac{\|(I-P(\theta))(\eta(\theta)-\eta(\theta_1))\|}{\|\eta(\theta)-\eta(\theta_1)\|^2}\left(\frac{(\theta-\theta_1)^t M(\theta)(\theta-\theta_1)}{\|\eta(\theta)-\eta(\theta_1)\|^2}\right)^{\alpha}.$$

Clearly,

$$K_{int,\alpha}(\theta,\theta_1)=\frac{1}{t(\theta,\theta_1)}\left(\frac{(\theta-\theta_1)^t M(\theta)(\theta-\theta_1)}{\|\eta(\theta)-\eta(\theta_1)\|^2}\right)^{\alpha},$$

where $t(\theta,\theta_1)$ is given by equation (1.3). It was shown in [15] that locally

$$K_{int}(\theta,\theta_1)=C_{int}(\theta,v)+\mathcal{O}(\|\theta-\theta_1\|),$$

where $v=\frac{\theta-\theta_1}{\|\theta-\theta_1\|}$. This shows that for every unit vector $v\in\mathbb{R}^m$,

$$\lim_{t\to 0}K_{int,\alpha}(\theta,\theta+tv)=C_{int}(\theta,v). \tag{4.1}$$

Eq. (4.1) establishes a relationship between $K_{int,\alpha}(\theta,\theta_1)$ and $C_{int}(\theta,v)$ when θ_1 approaches θ along the direction of a unit vector v. Note that $K_{int,\alpha}(\theta,\theta_1)$ requires appropriate selection of α. While no general recommendation regarding selection of α is provided in the literature, the choice of $\alpha=\frac{1}{2}$ is shown in [15] to be suitable in Examples 3.1 and 3.2.

In our opinion,

$$\sup_{\theta_1\in\Theta}K_{int,0}(\theta,\theta_1)=\frac{1}{\inf_{\theta_1\in\Theta}t(\theta,\theta_1)}=\frac{1}{d(\theta)}$$

is a natural, geometrically meaningful, extension of intrinsic curvature $C_{int}(\theta)$ given in Eq.(1.5). Recall that the radius of intrinsic curvature, $R_{int}(\theta)=\frac{1}{C_{int}(\theta)}$ is

$$\sup\{R:\ \|y-\eta(\theta)\|_W\le\|y-\eta(\theta_1)\|_W\}$$

if y satisfies

$$\|y-\eta(\theta)\|_W\le R\quad\text{and}\quad\|\theta-\theta_1\|<\delta,\ \text{with some }\delta(y). \tag{4.2}$$

Similarly, it follows from Theorem 1 in [9] that

$$d(\theta)=\sup\{R:\ \|y-\eta(\theta)\|_W\le\|y-\eta(\theta_1)\|_W\}$$

if y satisfies (4.2) and $\|y-\eta(\theta)\|_W\le R$.

Thus, $d(\theta)$ is a direct extension of $R_{int}(\theta)$ that removes the local condition, $\|\theta-\theta_1\|<\delta$, meaning that $d(\theta)\le R_{int}(\theta)$ or, equivalently,

$$K_{int}(\theta)\ge C_{int}(\theta).$$

Moreover, similarly to [15, p. 207],

$$\lim_{t\to 0}t(\theta,\theta+tv)=\frac{1}{C_{int}(\theta,v)},$$

and

$$\inf_v\lim_{t\to 0}t(\theta,\theta+tv)=\frac{1}{\sup_v C_{int}(\theta,v)}=\frac{1}{C_{int}(\theta)}=R_{int}(\theta).$$

5. Discussion and conclusion

A new intrinsic MoN is proposed in this paper as a natural extension of $C_{int}(\theta)$. It was shown that the radius of curvature, $d(\theta)$, together with associated quantities $t(\theta, \theta_1)$ and $d(\theta)$ (Eqs. (1.3) and (1.4)), lead to new Criteria 3 and 4 for a global minimizer of the SS. As illustrated in Examples 3.1–3.3, the new criteria work when Criteria 1 and 2 don't. Most importantly, Criteria 3 and 4 do not impose any restrictions on the parameter space Θ (such as compactness in Criterion 1) and require only optimization, but not other assumptions like connectedness of the level set $L(\theta)$ in Demidenko's Criterion 2, which could be difficult to verify. While global minimization over multi-parameter space may present some challenges, it is simplified in the presence of linear parameters, as illustrated in Examples 3.3 and 3.4. As follows from our previous work [9, 11], Criteria 3 and 4 offer sharp boundaries in the class of criteria that majorise $S(\theta)$ using either a function of θ or a constant, and do not impose any additional requirements. Moreover, as illustrated in Example 3.3, a hybrid of Criteria 3 and 4 is possible, where minimization of the function estimating $S(\theta)$ occurs over some subset of the parameter space rather than over the entire parameter space. This is typically the case when the model contains both linear an nonlinear parameters, like in Examples 3.3 and 3.4. In this case neither Criterion 4 nor Criterion 2 work since both d and $\bar{S}_{LU}$ are 0 when the model's linear parameter varies over $(0, \infty)$ or any other set for which 0 is an accumulation point. In this case either Criterion 3 or a hybrid criterion based on Criterion 3 could be applied.

An important practical problem to quantify the variance of the model's errors ensuring that the parameter estimation criterion readily identifies the global minimizer is beyond the scope of this paper. The solution to the problem presented in ([10]) in the case of normally distributed errors is based on the notions of the equidistant function $t(\theta, \theta_1)$ (Eq. (1.3)) and the radius of curvature $d(\theta)$ (Eq. (1.4)) discussed in this paper.

The new measure of nonlinearity introduced in this paper merits further exploration due to its geometric appeal coupled with its applicability to different aspects of nonlinear parameter estimation and inference.

References

[1] E. M. L. Beale, *Confidence regions in nonlinear estimation*, Journal of the Royal Statistical Society **22** (1960), 41–76.

[2] D. M. Bates D. G. Watts, *Relative curvature measures of nonlinearity (with discussion)*, Journal of the Royal Statistical Society: Series B **42** (1980), 1–25.

[3] G. Chavent, *A new sufficient condition for the well-posedness of non-linear least-square problems arising in identification and control*, in: A. Bensoussan, J. L. Lions (eds.), Analysis and Optimization of Systems, Lecture Notes in Control and Information Sciences **144**, Springer, Berlin, Heidelberg, 1990, pp. 452–463.

[4] E. Demidenko, *Is this the least squares estimate?*, Biometrika **87** (2000), 437–452.

[5] E. Demidenko, *Criteria for global minimum of sum of squares in nonlinear regression*, Computational Statistics and Data Analysis **51** (2006), 1739–1753.

[6] D. C. Hamilton, *Accounting for intrinsic nonlinearity in nolinear regression parameter inference regions*, Biometrika **73** (1986), 57–64.

[7] D. C. Hamilton, D. G. Watts and D. M. Bates, *Accounting for intrinsic nonlinearity in nonlinear regression parameter inference regions*, The Annals of Statistics **10** (1982), 386–393.

[8] L. Haines, *A note on the differential geometry of least squares estimation for nonlinear regression models*, South African Statistical Journal **28** (1994), 73–91.

[9] L. Khinkis and M. Crotzer, *A new approach for finding global minima in nonlinear least squares regression*, in: Proceedings of the American Statistical Association, Biopharmaceutical Section, Joint Statistical Meeting, 2008, pp. 2264–2271.

[10] L. Khinkis and M. Crotzer, *Application of new measures of nonlinearity to parameter estimation and simulations in individual pharmacokinetic analyses*, Journal of Pharmacokinetics and Pharmcodynamics **46** (2019), 43–52.

[11] L. Khinkis, M. Crotzer and A. Oprisan, *Sizing up the regions of unique minima in the least squares nonlinear regression*, Mathematics for Applications **7** (2018), 41–52.

[12] L. Michaelis and M. L. Menten, *Kinetics for intertase action*, Biochemische Zeitung **49** (1913), 333–369.

[13] A. Pázman, *Nonlinear Statistical Models*, Kluwer, Dordrecht, 1993.

[14] A. Pázman and L. Pronzato, *Optimum design accounting for the global nonlinear behavior of the model*, The Annals of Statistics **42** (2014), 1426–1451.

[15] L. Pronzato and A. Pázman, *Design of Experiments in Nonlinear Models*, Springer, 2013.

Leonid Khinkis, School of Mathematical Sciences, Rochester Institute of Technology, Rochester, NY, USA
e-mail: laksma@rit.edu