Gradient Projection Confidence Intervals in Nonlinear Regression

Jian-Shen Chen

We consider confidence intervals for parametric functions \( g(\theta) \) of the parameters \( \theta \) in a nonlinear regression model. It is known that the standard linear approximation (LA) confidence intervals can be quite poor. The alternative to the LA interval is the likelihood ratio (LR) interval, but to obtain the LR interval is computationally complex and expensive. We propose and investigate a gradient projection (GP) method to construct confidence intervals for \( g(\theta) \). In our view it is considerably simpler than the LR method, but will share its superior coverage properties. Conditions are identified under which the GP intervals are exact. The performance of the GP method in a variety of nonlinear regression problems is investigated via simulation study and for these problems the proposed method works quite well.

KEY WORDS: Least Squares; Linear Approximation Interval; Likelihood Ratio Interval; Linearizable; Exactness; Simulation Study.
I. INTRODUCTION

Confidence intervals are very basic and important in many areas of applications such as clinical trials, survival analysis, and toxicology studies. Two well-established approaches, the linear approximation (LA) method and the likelihood ratio (LR) method, can be employed to construct confidence intervals in a nonlinear regression model. For detailed discussions, see Chen (1991), Chen and Jennrich (1995a), Clarke (1987), Donaldson and Schnabel (1987), Gallant (1975), and Seber and Wild (1989). However, as noted, the LA method is less reliable and the LR method requires more computational effort. We would like to propose and investigate a gradient projection (GP) method which is considerably more reliable than the LA method and will be simpler and less expensive than the LR method.

We are interested in confidence intervals for a parametric function \( g(\theta) \) of the parameters \( \theta \) in the nonlinear regression model

\[
y = f(\theta) + \epsilon \sim N(0, \sigma^2 I)
\]  

(1.1)

Here \( y \) is an \( n \)-vector of observed responses, \( f(\theta) \) is a known vector valued function of an unknown parameter vector \( \theta \) that ranges over a \( p \)-dimensional parameter space \( \Theta \). Because our primary interest is in the performance of the proposed method, other than departures from normality, we have assumed that \( \epsilon \) is an unknown error vector that satisfies the standard normality assumption.

The least squares method for nonlinear regression models minimizes the residual sum of squares

\[
S(\theta) = ||y - f(\theta)||^2
\]  

(2.1)

where \( \cdot \) denotes the \( l_2 \)-norm. If the parameter space \( \Theta \) is compact, the least squares estimate exists and under normality assumption it is the maximum likelihood estimate of the parameters \( \theta \).

To see the concept of the GP method, let \( M \) be the range of \( f \). This is sometimes called the solution locus (Box and Lucas 1959) or response manifold because it is a \( p \)-dimensional differentiable manifold when \( \frac{df}{d\theta} \) is continuous and has full column rank for all \( \theta \) where \( \frac{df}{d\theta} \) is the Jacobian of \( f \) at \( \theta \). The conceptually and technically important first step, not always made in this area, is to view \( g \) as a function of \( f \) rather than a function of \( \theta \). More precisely we assume \( g \) is a real valued function defined on \( M \). Denote \( p = f(\theta) \) and define \( \hat{g} = g(p) \).

In Section 2 we first briefly discuss the background of the LA method and the LR method. Secondly we define the GP interval and compare it to the two intervals previously mentioned. In Section 3 two conditions are identified, which make the GP interval exact. Moreover, Theorem 3.4 set forth sufficient conditions for the exactness of the LR interval. In Section 4 the performance of the GP interval for a variety of nonlinear regression models is evaluated by simulation study. The results show that in our examples the observed coverage probabilities are close to the nominal level. Finally in the Appendix we give models and data sets used in our simulation studies.

II. GRADIENT PROJECTION INTERVALS

So far the most commonly used confidence intervals in nonlinear regression are the LA intervals. We will define these intervals based on the following result. It can be shown (see, Chen and Jennrich, 1995a) that the gradient of \( \hat{g} \) at \( \mu \)

\[
\nabla \hat{g}(\mu) = \frac{df}{d\theta} \left( \frac{df}{d\theta} \right)^{-1} \frac{d^2 f}{d\theta^2}
\]  

(2.1)

where \( \frac{df}{d\theta} \) is the Jacobian of \( g \) at \( \theta \). Let \( \hat{\theta} \) be the least squares estimate of \( \theta \). In terms of \( y \) the LA intervals with nominal level \( 1 - \alpha \) take the form

\[
\hat{g}(\mu) - t_{\alpha/2, n-p} \|\nabla \hat{g}(\mu)\| \sigma < \hat{g}(\mu) < \hat{g}(\mu) + t_{\alpha/2, n-p} \|\nabla \hat{g}(\mu)\| \sigma
\]  

(2.2)

where \( \hat{\mu} = f(\hat{\theta}) \), \( t_{\alpha/2, n-p} \) is the \( \alpha/2 \) upper quantile of Student's t-distribution with \( n-p \) degree of freedom, \( \nabla \hat{g} \) is the gradient of \( \hat{g} \) at \( \hat{\mu} \), and

\[
\hat{\sigma}^2 = \frac{||y - \hat{\mu}||^2}{n - p}
\]  

(2.3)

These are very easy to compute and are standard output from nonlinear least squares programs. They frequently work well, but occasionally behave badly (see, for example, Chen and Jennrich, 1995a; Donaldson and Schnabel, 1987) primarily by producing low coverage probabilities. The probability that the true value of \( g(\theta) \) falls within its confidence interval is called coverage probability, denoted by \( CP \).

The principle alternative to the LA intervals are the LR intervals. Let \( \hat{\theta}_c = f(\hat{\theta}_c) \) where \( \hat{\theta}_c \) is the nonlinear least squares estimate of \( \theta \) given \( \hat{g}(\mu) = c \). The LR intervals with nominal level \( 1 - \alpha \) are sets of the form

\[
\{ c : ||y - \hat{\mu}||^2 - ||y - \hat{\mu}||^2 < t_{\alpha/2, n-p}^2 \}
\]  

(2.4)

These require search from \( -\infty \) to \( \infty \). These need not be intervals or even bounded sets and can be difficult to compute. Their computation is expensive and tedious because it requires a
sequence of constrained least squares fits. Their advantage is that their coverage probabilities appear to be much better than those of the LA intervals. Actually the LR method, probably because of its complexity, is seldom used. There are variations of the LA method and of the LR method, but these are seldom used.

Let \( d = \nabla \hat{g}(\mu) \). This means the vector \( \nabla \hat{g}(\mu) \) scaled to have length one. The definition of the GP confidence interval for \( \hat{g}(\theta) \) with nominal level \( 1 - \alpha \) is

\[
\hat{g}(\mu - t_{\alpha} \hat{d}) < \hat{g}(\theta) < \hat{g}(\mu + t_{\alpha} \hat{d})
\]

where \( \hat{d} \) is the direction of the gradient of \( \hat{g} \) at \( \hat{\mu} \). This always provides an interval. The boundary problem may occur in finding the GP intervals. We will not investigate this here. To set the limits, one can form the pseudo data

\[
\hat{u} = \mu - t_{\alpha} \hat{d}
\]

and find the corresponding least squares estimates \( \hat{u}^* \) and \( \hat{u}^*_o \). Then the lower and upper limits of the GP intervals are \( \hat{g}(\hat{u}^*) \) and \( \hat{g}(\hat{u}^*_o) \). This is a simple straightforward procedure which requires very little modification to standard nonlinear regression programs.

If \( f \) and \( g \) are linear, the LA interval (2.2), the LR interval (2.4), and the GP interval (2.5) are identical and exact by which we mean they have coverage probabilities \( 1 - \alpha \). Moreover, these three intervals are invariant under reparametrization of \( f \) and \( g \). The LA interval and the GP interval are invariant because \( \hat{g}, \hat{\mu}, \) and \( \hat{d} \) are invariant under reparametrization of \( f \) and \( g \). The reason for the LR interval is because of invariance of the likelihood function under reparametrization.

III. PROPERTIES OF GRADIENT PROJECTION INTERVALS

We say that \( h \) is generalized linear function if

\[
h(y) = u(l^T y)
\]

for some function \( u \) and \( n \)-vector \( l \). The following theorem is proved by Chen and Jennrich (1995a).

**Theorem 3.1.** If \( h \) is a generalized linear function on \( \mathbb{R}^n \), \( q \in \mathbb{R}^n \), \( d = \nabla \hat{b}(g) \), and

\[
h(x) = h(q + xd)
\]

for all \( x \in \mathbb{R}^n \), then

\[
h(y) = \hat{h}(d^T (y - q))
\]

for all \( y \in \mathbb{R}^n \).

If \( M \) is a linear manifold, it follows from general linear model theory that \( \hat{\theta} \) and \( \hat{\sigma} \) are independent, and that

\[
\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p} \chi^2(n-p)
\]

We say that \( \hat{g} \) is linearizable if there is a strictly monotone transformation \( T \) such that \( T(\hat{g}) \) is linear. Let

\[
\hat{g}(x) = \hat{g}(x + \hat{d})
\]

It can be shown that if \( \hat{g} \) is linearizable, then \( \hat{g} \) is invertible (see, Chen and Jennrich, 1995a). Theorem 3.2. If \( \hat{g} \) is linearizable, and \( M \) is linear of dimension \( p \), then the GP interval is exact.

**Proof.** From Theorem 3.1

\[
CP = P(\hat{u}^* \hat{d} < \hat{g}(\theta) < \hat{u}^*_o \hat{d})
\]

where \( \hat{u}^* \) and \( \hat{u}^*_o \) are the projections of \( \hat{g}(\mu) \) and \( \hat{g}(\mu) \) onto \( M \) and \( M_{\mu_0} \), respectively. We obtain

\[
CP = P \left[ \frac{S(\hat{\theta}) - S(\hat{\theta})}{S(\hat{\theta})} < |p - \hat{\theta}|^2 \right]
\]

\[
= P \left[ \frac{|p - \hat{\theta}|^2 - |p - \hat{\theta}|^2}{|p - \hat{\theta}|^2} < |p - \hat{\theta}|^2 \right].
\]
Since $M_n \in M$, it follows from general linear model theory that the last expression in (3.8) is equal 1 – $\alpha$. The theorem is proved.

From Theorem 3.4 the linearity of $M_n$ and $M$ are basic assumptions. Thus one may want to use the measures of nonlinearity of $M$ and $M_n$ as a diagnostic tool for the LR interval. The result following is easy to prove that expresses the identical case between the GP interval and the LR interval. If $\hat{g}$ is linearizable, and $M$ is linear of dimension $p$, then the GP and the LR intervals are identical.

IV. SIMULATION RESULTS

The Monte Carlo method is used to simulate the 14 model function/data set combinations provided in the Appendix. The observed coverage probabilities of the nominal 95% confidence intervals constructed using the LA method, the GP method, and the LR method are tabulated in Table 1. Each simulation used 1000 generated data sets. The simulation estimates for the true coverage probabilities have standard errors of about 1%. The simulation results show that the GP method and the LR method provide very reliable confidence intervals since the observed coverage probabilities are within two simulation standard deviations of the nominal level more than 95% of the time. It can be seen that the observed coverage probabilities of the 95% confidence intervals constructed using the GP method are as close to the nominal level as those of the LR method. These statistically support our previous argument that the GP method is as good as the LR method for constructing confidence intervals for $g(\theta)$. The advantage of the GP method is that it is simple and inexpensive to produce the confidence intervals for any parametric function. Any existing computer codes implemented to solve the nonlinear least squares estimates can also be employed for the GP method.

On the other hand, comparing with the nominal 95% confidence level, the observed coverage probabilities of confidence intervals constructed using the LA method are poor for some of the parametric functions. For example, for the MMF model (Data set 12), the observed coverage probabilities for $\theta_j$ is as low as 83.2% by using the LA method. This means that the LA interval is much less accurate and reliable than the other two intervals. From the simulated results and observations, the LA method confidence intervals for some of the parametric functions could be extremely misleading. However, the LA method seems to work well for many of the examples. Therefore it would be fairly helpful to have a diagnostic tool available for detecting the low coverage probabilities of the LA intervals. See Chen and Jennrichi (1995a, 1995b) for detailed discussions.

<table>
<thead>
<tr>
<th>Models</th>
<th>$g(\theta)$</th>
<th>LA</th>
<th>GP</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMM</td>
<td>$\theta_1$</td>
<td>95.0</td>
<td>94.7</td>
<td>94.7</td>
</tr>
<tr>
<td></td>
<td>$\theta_2$</td>
<td>94.7</td>
<td>93.9</td>
<td>94.0</td>
</tr>
<tr>
<td>BOD</td>
<td>$\theta_1$</td>
<td>94.0</td>
<td>93.8</td>
<td>94.9</td>
</tr>
<tr>
<td></td>
<td>$\theta_2$</td>
<td>94.2</td>
<td>95.5</td>
<td>95.3</td>
</tr>
<tr>
<td>BOX</td>
<td>$\theta_1$</td>
<td>93.7</td>
<td>93.6</td>
<td>94.0</td>
</tr>
<tr>
<td></td>
<td>$\theta_2$</td>
<td>95.0</td>
<td>94.8</td>
<td>95.0</td>
</tr>
<tr>
<td>EAM</td>
<td>$\theta_1$</td>
<td>95.5</td>
<td>96.3</td>
<td>96.3</td>
</tr>
<tr>
<td></td>
<td>$\theta_2$</td>
<td>95.9</td>
<td>96.0</td>
<td>96.1</td>
</tr>
<tr>
<td>CGM</td>
<td>$\theta_1$</td>
<td>86.3</td>
<td>94.9</td>
<td>94.9</td>
</tr>
<tr>
<td></td>
<td>$\theta_2$</td>
<td>95.2</td>
<td>94.7</td>
<td>94.8</td>
</tr>
<tr>
<td>ARM</td>
<td>$\theta_1$</td>
<td>85.1</td>
<td>93.8</td>
<td>94.0</td>
</tr>
<tr>
<td></td>
<td>$\theta_2$</td>
<td>84.1</td>
<td>93.0</td>
<td>94.1</td>
</tr>
<tr>
<td></td>
<td>$\theta_3$</td>
<td>94.1</td>
<td>93.8</td>
<td>93.9</td>
</tr>
<tr>
<td>LOG</td>
<td>$\theta_1$</td>
<td>94.4</td>
<td>94.5</td>
<td>94.6</td>
</tr>
<tr>
<td></td>
<td>$\theta_2$</td>
<td>95.2</td>
<td>95.3</td>
<td>95.4</td>
</tr>
<tr>
<td></td>
<td>$\theta_3$</td>
<td>95.8</td>
<td>93.8</td>
<td>96.0</td>
</tr>
<tr>
<td>GOM</td>
<td>$\theta_1$</td>
<td>93.5</td>
<td>94.9</td>
<td>95.1</td>
</tr>
<tr>
<td></td>
<td>$\theta_2$</td>
<td>94.9</td>
<td>94.9</td>
<td>95.0</td>
</tr>
<tr>
<td></td>
<td>$\theta_3$</td>
<td>95.5</td>
<td>95.0</td>
<td>94.9</td>
</tr>
<tr>
<td>CTM</td>
<td>$\theta_1$</td>
<td>94.2</td>
<td>94.2</td>
<td>94.2</td>
</tr>
<tr>
<td></td>
<td>$\theta_2$</td>
<td>94.3</td>
<td>94.3</td>
<td>94.3</td>
</tr>
<tr>
<td></td>
<td>$\theta_3$</td>
<td>94.3</td>
<td>94.3</td>
<td>94.3</td>
</tr>
<tr>
<td>MRM</td>
<td>$\theta_1$</td>
<td>94.4</td>
<td>94.3</td>
<td>94.3</td>
</tr>
<tr>
<td></td>
<td>$\theta_2$</td>
<td>94.5</td>
<td>94.4</td>
<td>94.4</td>
</tr>
<tr>
<td></td>
<td>$\theta_3$</td>
<td>94.7</td>
<td>94.6</td>
<td>94.6</td>
</tr>
<tr>
<td>FQM</td>
<td>$\theta_1$</td>
<td>95.2</td>
<td>95.2</td>
<td>95.2</td>
</tr>
<tr>
<td></td>
<td>$\theta_2$</td>
<td>94.1</td>
<td>94.9</td>
<td>95.1</td>
</tr>
<tr>
<td></td>
<td>$\theta_3$</td>
<td>94.1</td>
<td>94.1</td>
<td>94.1</td>
</tr>
<tr>
<td>MMF</td>
<td>$\theta_1$</td>
<td>95.2</td>
<td>94.9</td>
<td>95.3</td>
</tr>
<tr>
<td></td>
<td>$\theta_2$</td>
<td>95.7</td>
<td>95.6</td>
<td>95.8</td>
</tr>
<tr>
<td></td>
<td>$\theta_3$</td>
<td>83.2</td>
<td>95.3</td>
<td>95.6</td>
</tr>
<tr>
<td></td>
<td>$\theta_4$</td>
<td>95.6</td>
<td>95.5</td>
<td>95.3</td>
</tr>
</tbody>
</table>

Table 1. Simulation Results.
Table 1 (continued).

<table>
<thead>
<tr>
<th>Data set 2 (Identifier HOD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model function: Michaelis-Menten model.</td>
</tr>
<tr>
<td>( f(x, \theta) = \frac{\theta_1 x}{\theta_2 + x} )</td>
</tr>
<tr>
<td>Data: ( n = 12, p = 2 ).</td>
</tr>
<tr>
<td>( x = (0.92, 0.02, 0.06, 0.06, 0.11, 0.01, 0.22, 0.22, 0.56, 0.56, 1.10, 1.10)^T ).</td>
</tr>
<tr>
<td>( y = (76.47, 97.167, 123.139, 150.152, 191.201, 207.200)^T ).</td>
</tr>
<tr>
<td>( \hat{\theta} = (212.6762, 0.06412)^T ).</td>
</tr>
<tr>
<td>( s^2 = 119.5449 ).</td>
</tr>
</tbody>
</table>

Data set 3 (Identifier BOX) |
Reference: Draper and Smith (1981), problem L, set 1, page 522 |
Model function: |
Data: \( n = 6, p = 2 \). |
\( x = (1.2, 3.4, 5.7)^T \). |
\( y = (8.3, 10.3, 19.0, 16.0, 15.6, 19.8)^T \). |
\( \hat{\theta} = (19.1426, 0.5311)^T \). |
\( s^2 = 6.4976 \). |

Data set 4 (Identifier BAM) |
Model function: |
Data: \( n = 7, p = 2 \). |
\( x = (1.2, 3.4, 5.6, 7)^T \). |
\( y = (82.112, 153.163, 176.192, 200)^T \). |
\( \hat{\theta} = (205.2720, 0.4306)^T \). |
\( s^2 = 50.4088 \). |

APPENDIX
Model Function/Data Set Combinations

- Data set 1 (Identifier MMM) |
Data set 5 (Identifier CGM)
Model function:
\[ f(x, \theta) = \theta_1 e^{\theta_2 x} \]
Data: \( n = 5, p = 2 \).
\[ x = (4, 10, 17, 22, 25)^T, \]
\[ y = (5, 20, 45, 66, 85)^T, \]
\[ \hat{\theta} = (0.4801, 1.6027)^T, \]
\[ s^2 = 2.4336. \]

Data set 7 (Identifier LOG)
Model function: Logistic model,
\[ f(x, \theta) = \frac{\theta_1}{1 + e^{-\theta_2 x}} \]
Data: \( n = 9, p = 3 \).
\[ x = (9, 14, 21, 28, 42, 57, 63, 70, 79)^T, \]
\[ y = (8.93, 10.80, 18.59, 22.33, 39.35, 56.11, 61.73, 64.62, 67.08)^T, \]
\[ \hat{\theta} = (188.6147, -193.1564, -0.006247)^T, \]
\[ s^2 = 10.0296. \]

Data set 8 (Identifier GOM)
Model function: Gompertz growth model,
\[ f(x, \theta) = \theta_1 e^{-e^{\theta_2 x}} \]
Data: \( n = 9, p = 3 \).
\[ x = (9, 14, 21, 28, 42, 57, 63, 70, 79)^T, \]
\[ y = (8.93, 10.80, 18.59, 22.33, 39.35, 56.11, 61.73, 64.62, 67.08)^T, \]
\[ \hat{\theta} = (72.4622, 13.7093, -0.06736)^T, \]
\[ s^2 = 1.3428. \]

Data set 9 (Identifier CTM)
Model function: Asymptotic regression model,
\[ f(x, \theta) = \theta_1 + \theta_2 e^{\theta_3 x} \]
Data: \( n = 9, p = 3 \).
\[ x = (9, 14, 21, 28, 42, 57, 63, 70, 79)^T, \]
\[ y = (8.93, 10.80, 18.59, 22.33, 39.35, 56.11, 61.73, 64.62, 67.08)^T, \]
\[ \hat{\theta} = (82.8323, 1.2237, 0.03707)^T, \]
\[ s^2 = 0.8726. \]
Model function:
\[ f(x, \theta) = e^{x_1} \]  
Data: \( n = 10, p = 3, \) 
\[ x = (-4, -3, -2, -1, 0, 1, 2, 3, 4, 5)^T, \]  
\[ y = (-3.4000, -2.6923, -1.0000, -0.03348, 0.0857, 0.6667, 1.1875, 1.6538, 3.3333, 1.0000)^T, \]  
\[ \hat{\theta} = (0.4889, 3.0467, -0.1008)^T, \]  
\[ s^2 = 0.07069. \]  

Data set 12 (Identifier MMF)
Reference: Ratkowsky (1983), data set 1\(^{st}\), page 88.
Model function: Morgan-Mercer-Flodin growth model,
\[ f(x, \theta) = e^{\theta_3} + e^{\theta_4 x} \]  
Data: \( n = 9, p = 4, \) 
\[ x = (9, 14, 21, 28, 42, 57, 63, 70, 79)^T, \]  
\[ y = (8.93, 10.80, 18.59, 22.33, 39.35, 56.11, 61.73, 64.62, 67.08)^T, \]  
\[ \hat{\theta} = (80.9392, 8.8945, 4.9757, 7.2575, 2.8280)^T, \]  
\[ s^2 = 2.7114. \]  

Data set 13 (Identifier FMM)
Reference: Ratkowsky (1983), data set 1\(^{st}\), page 88.
Model function: Morgan-Mercer-Flodin growth model,
\[ f(x, \theta) = \frac{e^{\theta_3} + e^{\theta_4 x}}{e^{\theta_3} + x^{\theta_4}} \]  
Data: \( n = 9, p = 4, \) 
\[ x = (9, 14, 21, 28, 42, 57, 63, 70, 79)^T, \]  
\[ y = (8.93, 10.80, 18.59, 22.33, 39.35, 56.11, 61.73, 64.62, 67.08)^T, \]  
\[ \hat{\theta} = (80.9392, 8.8945, 4.9757, 7.2575, 2.8280)^T, \]  
\[ s^2 = 2.7114. \]  

Data set 14 (Identifier WBT)
Reference: Ratkowsky (1983), data set 1\(^{st}\), page 88.
Model function: Weibull type model,
\[ f(x, \theta) = \theta_1 - \theta_2 e^{-\theta_3 x^\theta_4} \]  
Data: \( n = 9, p = 4, \) 
\[ x = (9, 14, 21, 28, 42, 57, 63, 70, 79)^T, \]  
\[ y = (8.93, 10.80, 18.59, 22.33, 39.35, 56.11, 61.73, 64.62, 67.08)^T, \]  
\[ \hat{\theta} = (80.9392, 8.8945, 4.9757, 7.2575, 2.8280)^T, \]  
\[ s^2 = 1.6752. \]
REFERENCES


