Improved Confidence Estimation in Noninear Regression Models Using Higher-Order Asymptotics

Thomas J. Harris[1] and P. James McLellan
Department of Chemical Engineering, Queen's University, Kingston, ON, K7L 3N6

February 19, 2009

[1] Author to whom correspondence should be addressed. E-mail address: tom.harris@chee.queensu.ca

Abstract

Likelihood-based approaches, including profile-likelihood, signed- likelihood, sample deviance and profile-t techniques improve inference for nonlinear regression over linearization-based asymptotic approaches. While providing exact likelihood regions, the significance level is approximate. This paper provides a review of the techniques based on higher-order asymptotics that have been proposed to provide corrections to the nominal significance level in likelihood-based inference approaches in the case of nonlinear regression with homoscedastic error variance. A new approach for computing higher-order asymptotic corrections for significance level as a constrained optimization problem is proposed for scalar-valued functions of parameters. Alternative expressions for higher-order asymptotic corrections are developed that provide more direct comparison between corrections for coordinate parameters and scalar-valued functions of these parameters. Stable and efficient computational approaches are discussed, and the techniques are demonstrated using a sequence of examples of increasing complexity.

Keywords: nonlinear regression, confidence intervals, maximum likelihood, parameter estimation, profiling, generalized profiling, higher-order asymptotics

1 Introduction

It has been twenty years since the publication by Donaldson and Schabel of ”Computational Experience with Confidence Regions and Confidence Intervals for Nonlinear Least Squares” (Donaldson and Schnabel (1987)) in this journal . The results from their study, and those of others including Bates and Watts (1988), indicate that confidence intervals and regions obtained using the asymptotic limiting distribution for the parameter estimates can be misleading for parameter estimation problems that arise routinely in the course of scientific and engineering research and development. A number of diagnostic measures have been proposed to provide guidance as to when the standard confidence intervals may be misleading. These measures can be broadly characterized as: i) those focussing on nonlinearity of the model formulation and parameterization, and ii) those focussing on warping or distortion of the likelihood function, which in turn depends on the model being estimated. Both approaches measure departures from asymptotic descriptions based on underlying linear theory.

In the case of the model formulation and parameterization, the asymptotic inference results based on linear theory rely on a tangent plane approximation to the expectation surface of model predictions, and a uniform coordinatization on that plane implied by the first-order Taylor series approximation. In the case of the likelihood function, departures are measured from a quadratic approximation to the likelihood function. Global measures, such as those that measure the maximum nonlinearity of the model function over all directions leading to departures from the linear approximation, or that measure the average nonlinearity, generally indicate when there may be problems with inference based on asymptotic results. However, significant maximum or average nonlinearity in the model function does not always translate to problems with the asymptotic inference results, both for estimated parameters and model predictions, and for other functions of the estimated parameters. The nonlinearity measures proposed in the literature generally provide limited guidance to the directions in which the departures due to nonlinearity are important.

Most of the curvature-based diagnostic measures do not not provide constructive approaches for forming reliable confidence intervals and confidence regions. Hamilton et al. (1982) use a combination of re-parameterization to reduce parameter-effects curvature, and a quadratic correction to reduce intrinsic curvature. However, intrinsic curvature, which is usually dictated by the form of the model and the experimental design, is usually not as significant a problem in nonlinear regression. Often, the intrinsic curvature is fixed from the experimental design used and the form of the model specified, and it is up to re-parameterization to reduce the more substantial effects of parameter-effects curvture (Ratkowsky (1983),Bates and Watts (1988)). Construction of more reliable inference regions generally requires numerically intensive methods, such as re-sampling or use of likelihood ratio methods. The focus in this paper is on recently developed adaptations to likelihood-based methods.

Likelihood-based methods for improved inference include profile-likelihood, signed-likelihood, sample deviance, and profile-t techniques. These methods provide exact descriptions of the likelihood region or intervals, however the nominal significance level associated with these regions is based on the asymptotic distribution of the likelihood ratio, which under suitable regularity conditions, is Chi-squared with an appropriately chosen degrees of freedom. The profile-t method assumes that the limiting distribution for a function of the parameters is given by the Student's t distribution. Many studies (e.g., Bates and Watts (1988), Chen and Jennrich (1996)) have shown that the coverage probabilities obtained from the likelihood approach are superior to those obtained from asymptotic theory. The vast majority of studies have focussed on nonlinear regression problems where the errors are independently and identically distributed (IID) Normal. There are many nonlinear regression problems that have error structures considerably more sophisticated than this, including multi-response estimation problems, problems with heteroscedastic errors such as those encountered in econometrics (e.g., ARCH or GARCH models), noise from non-Normal distributions, and situations in which the regressors have uncertainty (errors-in-variables models).

In the past twenty years, there has been considerable theoretical development of methods that provide correction factors to the limiting distribution of the likelihood ratio. Most of these were developed out of the seminal work of Barndorff-Nielsen and Cox (1984). Reviews are provided in Reid (1996), Skovgaard (2001), and Strawderman (2000). Most results have focussed on providing more reliable confidence estimation for scalar functions of interest, including coordinate parameters (one of the individual parameters in the model), nonlinear functions of the model parameters such as predictions, or more complicated functions of the parameters which may include contributions from the noise variance. Examples of the application of higher-order asymptotic corrections are discussed in Bellio et al. (2000) and Brazzale et al. (2007). Bellio and Brazzale (2003) describe a computational platform for the implementation of higher-order asymptotics in nonlinear regression problems having heteroscedastic errors using R.

The paper begins by providing a review of the techniques that have been proposed and how they can be applied to functions of parameters. A new approach for computing these corrections as constrained optimization problems, building on an approach proposed by Fraser, Wong and Wu (1999) and Brazzale et al. (2007), is presented. Alternative expressions for the higher-order asymptotic corrections are developed that provide a more transparent comparison between corrections for coordinate parameters and those for scalar-valued functions of the coordinate parameters. A comprehensive discussion of numerically efficient and stable computational approaches is provided, and the paper ends by presenting several examples of differing complexity in which the efficacy of Bates and Watts (1988).

2 Preliminaries

2.1 Model Description

Consider a general model of the form:

$$y_t = f(\boldsymbol{x}_t, \boldsymbol{\theta}) + \epsilon_t \qquad (t = 1, 2, ...n) \tag{1}$$

where the function $f(\boldsymbol{x}_t, \boldsymbol{\theta})$ is the expected value of the response variable y_t, $\boldsymbol{x}_t$ is a vector of the levels of m independent variables, $\boldsymbol{x}_t = (x_{1t}, x_{2t}, \ldots, x_{mt})^T$, $\boldsymbol{\theta}$ is a vector of p parameters $(\theta_1, \theta_2, \ldots, \theta_p)^T$, and ϵ_t is the additive random error term associated with y_t. The independent variables $\boldsymbol{x}_t$ are also referred to as input variables or manipulated variables. y_t is also called an output variable or a controlled variable. Equation(1) can be written in terms of a vector of n observed response values $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$:

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{f}(\boldsymbol{\theta}) + \boldsymbol{\epsilon} \qquad (2) \\ &= \begin{pmatrix} f(\boldsymbol{x}_1, \boldsymbol{\theta}) & \cdots & f(\boldsymbol{x}_n, \boldsymbol{\theta}) \end{pmatrix}^T + \boldsymbol{\epsilon} \end{aligned}$$

where $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of ϵ_t values.

In most applications of nonlinear regression it is assumed that the stochastic component, ϵ_t, is IID Normal. The applications of higher-order statistics allow considerably more flexibility in the specification of the underlying distribution. The requirement of independence is required , but the requirement that the underlying distribution be Normal can be considerably relaxed (Fraser, Wong and Wu (1999), Brazzale et al. (2007)). The use of a more flexible structure for the stochastic component and the use of higher-order corrections also means that one must move away from the nomenclature of sums of squares and use the notation of likelihood functions and sample information matrices. However, in this paper we will focus on the case in which the noise is IID Normal.

2.2 Inference for Parameters

To determine the statistical properties of the estimated parameters, let $\boldsymbol{\theta}^*$ denote the true value of the parameters, and let $\boldsymbol{\theta}$ denote a vector of arbitrary values. The least squares estimates, $\hat{\boldsymbol{\theta}}$, which coincide with the maximum likelihood estimates in this problem formulation, are those that are obtained from the solution to

$$\hat{\boldsymbol{\theta}} = arg \min_{\boldsymbol{\theta}} S(\boldsymbol{\theta}) \tag{3}$$

where

$$S(\boldsymbol{\theta}) = \sum_{t=1}^{n} \left(y_t - f(\boldsymbol{x_t}, \boldsymbol{\theta})\right)^2 \tag{4}$$

The unbiased estimate of the noise variance is given by

$$s^2 = \frac{S(\hat{\boldsymbol{\theta}})}{n-p} \tag{5}$$

Note that the maximum likelihood estimate of the noise variance, $\hat{\sigma}^2$, is obtained by dividing the residual sum of squares $S(\hat{\boldsymbol{\theta}}))$ by n, and not $n-p$.

The asymptotic distribution of the parameter estimates can be obtained by linearization. Define

$$\boldsymbol{F_*} = \frac{\partial \boldsymbol{f(\theta)}}{\partial \boldsymbol{\theta^T}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \qquad \hat{\boldsymbol{F}} = \frac{\partial \boldsymbol{f(\theta)}}{\partial \boldsymbol{\theta^T}} |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$
$$\boldsymbol{V_*} = \boldsymbol{F_*}^T \boldsymbol{F_*} \qquad \hat{\boldsymbol{V}} = \hat{\boldsymbol{F}}^T \hat{\boldsymbol{F}}$$

Theorem 1 (Seber and Wild (1989); Gallant (1987)): Let $\boldsymbol{f(\theta)}$ be a twice-differentiable mapping in the domain of interest. Given $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ and appropriate regularity conditions, then asymptotically

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{V}_*^{-1}) \tag{6}$$

$$\frac{[S(\boldsymbol{\theta}^*) - S(\hat{\boldsymbol{\theta}})]/p}{S(\hat{\boldsymbol{\theta}})/(n-p)} \sim F_{p,n-p} \tag{7}$$

where $N(\boldsymbol{0}, \sigma^2 \boldsymbol{V_*^{-1}})$ denotes a multivariate Normal distribution with mean vector $\boldsymbol{0}$ and variance-covariance matrix $\sigma^2 \boldsymbol{V_*^{-1}}$. $F_{p,n-p}$ denotes the ratio of two appropriately scaled independent χ^2 variables, with p and $n-p$ degrees of freedom respectively. On estimating $\boldsymbol{F_*}$ by $\hat{\boldsymbol{F}}$ and $\boldsymbol{V}_*^{-1}$ by $\hat{\boldsymbol{V}}^{-1}$ an approximate $100(1-\alpha)\%$ confidence interval for parameter θ_i^* is given by

$$\hat{\theta}_i \pm t_{n-p,\alpha/2} se(\hat{\theta}_i) \tag{8}$$

where the standard error is

$$se(\hat{\theta}_i) = \sqrt{s^2 [\hat{\boldsymbol{V}}^{-1}]_{ii}} \tag{9}$$

$t_{n-p,\alpha/2}$ denotes the critical value from the Student's t distribution with $n-p$ degrees of freedom. The justifications for replacing $\boldsymbol{F_*}$ by $\hat{\boldsymbol{F}}$ and $\boldsymbol{V}_*^{-1}$ by $\hat{\boldsymbol{V}}^{-1}$are outlined in Seber and Wild (1989) and Gallant (1987).

2.3 Inference for Scalar Functions of the Parameters

Consider a scalar-valued function of the parameters of the form $g(\boldsymbol{x}, \boldsymbol{\theta})$. Define the parametric sensitivity of g as:

$$\boldsymbol{g}_{\boldsymbol{*}} = \frac{\partial g(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$$
$$\hat{\boldsymbol{g}} = \frac{\partial g(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

and denote the estimate of g as:

$$\hat{g} = g(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$$

Theorem 2 (Seber and Wild Seber and Wild (1989)): In addition to the requirements in Theorem 1, let $g(\boldsymbol{x}, \boldsymbol{\theta})$ be twice-differentiable with respect to $\boldsymbol{\theta}$ in the region of interest. Asymptotically,

$$g(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) - g(\boldsymbol{x}, \boldsymbol{\theta}^*) \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{g}_{\boldsymbol{*}} \boldsymbol{V}_*^{-1} \boldsymbol{g}_{\boldsymbol{*}}{}^T) \tag{10}$$

On estimating $\boldsymbol{F}_{\boldsymbol{*}}$ by $\hat{\boldsymbol{F}}$, $\boldsymbol{g}_{\boldsymbol{*}}$ by $\hat{\boldsymbol{g}}$ and $\boldsymbol{V}_*^{-1}$ by $\hat{\boldsymbol{V}}^{-1}$, an approximate $100(1-\alpha)\%$ confidence interval for $g(\boldsymbol{x}, \boldsymbol{\theta}^*)$ is given by

$$g(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) \pm t_{n-p,\alpha/2} se(\hat{g}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})) \tag{11}$$

where

$$se(\hat{g}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})) = \sqrt{s^2 \hat{\boldsymbol{g}}^T \hat{\boldsymbol{V}}^{-1} \hat{\boldsymbol{g}}} \tag{12}$$

3 Generalized Profiling

Estimation of parameters in algebraic or differential equation models is often accomplished using a modest amount of data, particularly if the experiments are expensive to conduct. Even in system identification, where one is identifying transfer function models from dynamic data, it is often argued that the linearization-based approach for inference is justified on the basis that the linear inference results are the asymptotic result as $n \to \infty$, and it is common to deal with large amounts of data. However, there has been little evidence to support these claims, and the amount of data which constitutes a "large amount" is seldom quantified. Profiling (Bates and Watts (1988); Chen (1991); Lam and Watts (1991); Chen and Jennrich (1996); Quinn et al. (1999, 2000)) is a graphical means by which to display inference results for parameters, and functions of parameters, of proposed models. Bates and Watts (1988) developed profiling specifically to summarize inferential results for parameters of nonlinear regression models. Chen and Jennrich (1996) developed the theory of profiling in terms of likelihood ratios and constrained optimization. Chen and Jennrich's formulation of the profiling algorithm is very general and can be used with many classes of models, including time series models (Quinn et al. (2005)). Furthermore, the constrained optimization approach readily admits the problem of computing inference results for functions of parameters (Clarke (1987); Chen and Jennrich (1996); Quinn et al. (1999, 2000)), so long as these functions are twice-differentiable in the domain of interest. An important feature of profiling is that any restrictions implied by the constraints are satisfied. The terms profiling and generalized profiling are often used interchangeably. The discussion that follows provides an overview of profiling for scalar-valued functions $g(\boldsymbol{x}, \boldsymbol{\theta})$, however it should be noted that profiling of parameters represents the case in which the function g is the identity map, i.e., $g(\boldsymbol{x}, \boldsymbol{\theta}) = \theta_i$.

3.1 Profiling Scalar-Valued Functions

Profiling of functions, whether scalar-valued or vector-valued, is derived from the asymptotic distribution of the likelihood function $L(\boldsymbol{\theta})$. Note that the likelihood function depends implicitly on the values of the regressors in the dataset, however these are fixed in the inference analysis for any estimation problem. A distinction is made for $g(\boldsymbol{x}, \boldsymbol{\theta})$, because there will be instances in which we are interested in a function of the parameters that is also a function of different values of the regressors, such as a prediction at conditions other than those in the experiment dataset. In this section, the case of profiling a scalar-valued function is discussed. Extensions to vector-valued functions, i.e., multiple constraints, are discussed in a subsequent section. Note also that the profiling technique can handle instances where the function also depends on the noise variance, i.e., $g(\boldsymbol{x}, \boldsymbol{\theta}, \sigma^2)$. For the remainder of this section however, we will consider $g(\boldsymbol{x}, \boldsymbol{\theta})$.

As will been seen in a later section, the logarithm of the likelihood function is in many cases proportional to $S(\boldsymbol{\theta})$ defined in Eq(4). This is the case for the current paper given the assumption of IID Normal noise random variables.

A nominal $100(1-\alpha)\%$ likelihood interval for $g(\boldsymbol{x}, \boldsymbol{\theta})$ is the set of all values of $g(\boldsymbol{x}, \boldsymbol{\theta})$ which are plausible given the available data. From standard asymptotic arguments (Cox and Hinkley (1974)),

$$LI\left(g(\boldsymbol{x}, \boldsymbol{\theta})\right) = \left\{ g(\boldsymbol{x}, \boldsymbol{\theta}) : -2 \ln \left(\frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})} \right) \leq \chi^2_{1,\alpha} \right\} \tag{13}$$

where $LI(g(\boldsymbol{x}, \boldsymbol{\theta}))$ is the **likelihood interval** for $g(\boldsymbol{x}, \boldsymbol{\theta})$ and $\chi^2_{1,\alpha}$ is the upper α quantile for the χ^2 distribution with 1 degree of freedom, $\hat{\boldsymbol{\theta}}$ is the vector of maximum likelihood parameter estimates, and $\boldsymbol{\theta}$ is any allowable vector of parameter values (Chen and Jennrich (1996)). Note that the likelihood ratio statistic *asymptotically* follows the χ^2 distribution Cox and Hinkley (1974), except in special cases where it is exact. Note also that the likelihood interval for $g(\boldsymbol{x}, \boldsymbol{\theta})$ may be disjoint (Bates and Watts (1988)). It is also important to note that the likelihood functions in Eq(13) are *concentrated* likelihood functions in which the maximum likelihood estimate of the noise variance ($\hat{\sigma}^2 = S(\hat{\boldsymbol{\theta}})/n$) has been substituted into the likelihood function.

The profile functions defined by Chen and Jennrich (1996) and Bates and Watts (1988) make use of the conditional estimate of the parameters. Define $\tilde{\boldsymbol{\theta}}$ as the maximum likelihood estimate of the parameters subject to the constraint $g(\boldsymbol{x}, \boldsymbol{\theta}) = c$ where c is a constant value:

$$\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \tag{14}$$

$$\textit{subject to } g(\boldsymbol{x}, \boldsymbol{\theta}) = c$$

Chen and Jennrich define the likelihood profile as the function $L(\tilde{\boldsymbol{\theta}})$ as a function of c (Chen and Jennrich (1996)). The *signed rood deviance*, or *signed likelihood*, is now defined as the statistic

$$r = r(g(\boldsymbol{x}, \tilde{\boldsymbol{\theta}})) = sign(g(\boldsymbol{x}, \tilde{\boldsymbol{\theta}}) - g(\boldsymbol{x}, \hat{\boldsymbol{\theta}}))\sqrt{-2\ln\left(\frac{L(\tilde{\boldsymbol{\theta}})}{L(\hat{\boldsymbol{\theta}})}\right)} \tag{15}$$

This expression for r is general in that it may be used to make inferences about functions of parameters of any model so long as an expression for the likelihood function may be found. Geometric interpretations of generalized profiling are given in Quinn et al. (1999) and Quinn et al. (2000).

Expressed in terms of the signed likelihood, a $100(1-\alpha)\%$ likelihood interval for $g(\boldsymbol{x}, \boldsymbol{\theta})$ is the set of all values c for which

$$-z_{\alpha/2} \leq r \leq z_{\alpha/2} \tag{16}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile for the standard Normal distribution.

A profile-r plot is a plot of $r(g(\boldsymbol{x}, \tilde{\boldsymbol{\theta}}))$ versus c for a range of values of c. The limits of the likelihood interval for $g(\boldsymbol{x}, \boldsymbol{\theta^*})$ can be read from the profile-r plot by finding those values of $g(\boldsymbol{x}, \tilde{\boldsymbol{\theta}})$ which define the points on the profile at $r = \pm z_{\alpha/2}$. (Jennrich and Chen refer to the profile-r plot as a profile-z plot, for reasons that will be apparent in the next few paragraphs.)

Often, it is of interest to judge the relative nonlinearity of a parameter, or function of parameters, so as to know how reliable the linearization-based inference results will be. A reference line, $\delta_r(g(\boldsymbol{x}, \tilde{\boldsymbol{\theta}}))$ versus

$g(\boldsymbol{x}, \tilde{\boldsymbol{\theta}}) = c$ for a range of values of c, is typically included on profile-r plots, where

$$\delta_r(g(\boldsymbol{x}, \tilde{\boldsymbol{\theta}})) = \frac{g(\boldsymbol{x}, \tilde{\boldsymbol{\theta}}) - g(\boldsymbol{x}, \hat{\boldsymbol{\theta}})}{se_r(g(\boldsymbol{x}, \hat{\boldsymbol{\theta}}))} \tag{17}$$

This reference line can be used to obtain the linearization-based confidence intervals for $g(\boldsymbol{x}, \boldsymbol{\theta}^*)$, and to judge the departure from the linear approximation (Chen (1991), Chen and Jennrich (1996)). The reference line is tangent to the profile r-curve at the maximum likelihood estimate. It is important to note that for comparison purposes, the standard error is computed using the maximum likelihood-estimate of the residual variance in the case of nonlinear regression with IID Normal noise:

$$\begin{aligned} se_r(g(\boldsymbol{x}, \hat{\boldsymbol{\theta}})) &= \sqrt{\hat{\sigma}^2 \hat{\boldsymbol{g}}^T \hat{\boldsymbol{V}}^{-1} \hat{\boldsymbol{g}}} \\ \hat{\sigma}^2 &= \frac{S(\hat{\boldsymbol{\theta}})}{n} \end{aligned} \tag{18}$$

While the shape of the likelihood region defined by Eq(16) is exact, the confidence level is nominal, and the probability that the function of interest lies in the aforementioned region may differ from $1 - \alpha$. Discrepancies arise not from the nonlinearity itself, but from using the asymptotic distribution for the likelihood ratio.

Bates and Watts defined the *profile-t function*, τ, which is is the signed function:

$$\tau = \tau(g(\boldsymbol{x}, \tilde{\boldsymbol{\theta}})) = sign(g(\boldsymbol{x}, \tilde{\boldsymbol{\theta}}) - g(\boldsymbol{x}, \hat{\boldsymbol{\theta}}))\sqrt{\frac{S(\tilde{\boldsymbol{\theta}}) - S(\hat{\boldsymbol{\theta}})}{s^2}} \tag{19}$$

Approximate limits of the likelihood interval for $g(\boldsymbol{x}, \boldsymbol{\theta}^*)$ can be read from the profile-t plot by finding those values of $g(\boldsymbol{x}, \tilde{\boldsymbol{\theta}})$ which define the points on the profile at $\tau = \pm t_{n-p,\alpha/2}$, where $t_{n-p,\alpha/2}$ is $\alpha/2$ quantile for the Student's t distribution with $n - p$ degrees of freedom (Cook and Weisberg (1990)). In this instance $\tilde{\boldsymbol{\theta}}$ is obtained by minimizing the sums of squares function subject to the constraint indicated in Eq(14). A reference line for assessing the nonlinear effects is constructed using $se(g(\boldsymbol{x}, \hat{\boldsymbol{\theta}}))$ instead of $se_r(g(\boldsymbol{x}, \hat{\boldsymbol{\theta}}))$ in Eq(18). Examples are given in Bates and Watts (1988), Lam and Watts (1991), Quinn et al. (1999), Quinn et al. (2000), Quinn et al. (2005), and Watts (1994). The use of the critical points from the Student's t distribution, rather than from the standard Normal distribution, is a pragmatic approach that recognizes the estimation of the noise variance.

4 Higher Order Asymptotics

Under the assumption of IID Normal noise, the likelihood intervals calculated from the profile-t function for models that are linear in the parameters are exact, however those from the profile-r function are only asymptotically correct. To see this, note that for models that are linear in the parameters, $\frac{\theta_i - \hat{\theta}_i}{s_{\hat{\theta}_i}} \sim t_{n-p}$, and:

$$\frac{1}{n-p} \frac{S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}})}{s^2} \sim \frac{\chi^2_p}{\chi^2_{n-p}} \tag{20}$$

Dividing the numerator of Eq(20) by p and the denominator by $n-p$ produces a quantity that is distributed as $F_{p,n-p}$ These are exact results, based on the distribution of $\hat{\boldsymbol{\theta}}$ for the linear case. However, for the likelihood ratio,

$$-2\ln\left(\frac{L(\tilde{\boldsymbol{\theta}})}{L(\hat{\boldsymbol{\theta}})}\right) = n\ln\left(\frac{S(\boldsymbol{\theta})}{S(\hat{\boldsymbol{\theta}})}\right) = n\ln\left(1 + \frac{S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}})}{S(\hat{\boldsymbol{\theta}})}\right)$$

which can be simplified for small differences in sums of squares to

$$\frac{S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}})}{S(\hat{\boldsymbol{\theta}})/n} = \frac{S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}})}{s^2} \frac{n}{(n-p)} \tag{21}$$

which *approximately* has a χ^2_p distribution under likelihood theory.

Comparing the likelihood results in Eq(21) and the Normal theory results in Eq(20) highlights the fact that there is an approximation $\chi^2_{n-p}/n \rightarrow 1$. The net result is that the likelihood ratio intervals formed are smaller than those obtained using exact Normal theory, even in the case of linear models, due to the fact that the uncertainty in the estimate of the residual variance has not been accounted for. Asymptotically, the results of both approaches are the same. Note that if the noise variance is known, and not estimated, then both methods give the same results for finite samples.

Considerable research has been undertaken to correct the asymptotic distribution of the likelihood ratio for the case of finite samples. These corrections are applicable to both linear and nonlinear problems. In this paper, we investigate the class of corrections pioneered by the work of Fraser, Reid and co-workers (Fraser, Reid and Wu (1999); Fraser, Wong and Wu (1999); Brazzale et al. (2007)). The corrections are of the form

$$r^* = r - \frac{1}{r} ln(\frac{r}{q}) \sim N(0,1) \tag{22}$$

where q is the higher-order correction, and r^* is the corrected signed likelihood. In the approach developed in the aforementioned references, q is derived from a score pivot and the nusiance parameters. The correction ensures that the confidence level is $1-\alpha+O(n^{-3/2})$ as opposed to $1-\alpha+O(n^{-1/2})$ for the uncorrected likelihood, where the nominal confidence level is $1-\alpha$ (Reid (1996), Fraser, Wong and Wu (1999)). It is interesting to note that the correction factors are asymptotic corrections to the Normal approximation. It might seem more reasonable to make corrections to the Student's t distribution since this correctly accounts for small sample effects for linear models, but this is not the case.

Before presenting the formulae for the correction factors, a very simple example will be used to give some insight into the nature of the correction.

4.1 Illustrative Example

Consider the non-linear regression problem for the model specified in Eq(2), in which the noise is IID Normal, and the model contains a *single* parameter θ. Using the framework presented in Fraser, Reid and Wu (1999) and Brazzale et al. (2007), the correction factor q is given by:

$$q = \frac{\hat{\sigma}}{\tilde{\sigma}} \frac{\hat{\boldsymbol{F}}^T \tilde{\boldsymbol{\varepsilon}}}{\sqrt{\hat{\boldsymbol{F}}^T \hat{\boldsymbol{F}}}} \tag{23}$$

where $\tilde{\boldsymbol{\varepsilon}}$ is the vector of scaled residuals with elements:

$$\tilde{\varepsilon}_i = \frac{\tilde{e}_i}{\tilde{\sigma}} \tag{24}$$

$\tilde{\sigma}^2$ is the mean square error of the residuals associated with the conditional maximum likelihood estimation ($\tilde{\sigma}^2 = S(\tilde{\theta})/(n)$), and $\tilde{e}_i$ are the residuals associated with the conditional problem in which θ is set to some particular value, $\tilde{\theta}$. $\hat{\sigma}^2$ is the maximum likelihood estimate estimate of the noise variance computed again for the unconditional case (here, $\hat{\sigma}^2 = S(\hat{\theta})/n$). Note that in this example, the Jacobian $\boldsymbol{F}$ of the vector $\boldsymbol{f}$ of model predictions is in fact a vector of sensitivities of the model predictions with respect to the lone parameter.

In the case of a single parameter, it isn't necessary to solve the constrained optimization problem in order to determine the profile r-function, since there is only one parameter. However, it is still necessary to solve the unconstrained problem to obtain the maximum likelihood estimates of the parameter $\hat{\theta}$ and the noise variance $\hat{\sigma}^2$. Note from Eq(23) that the correction factor depends in a straightforward way on the vector of sensitivities of the expectation function evaluated at the maximum likelihood estimates. This is generally not the case. Now consider two special cases.

4.1.1 Linear Model

For linear models containing a single parameter θ, it is readily established that

$$q = (\tilde{\theta} - \hat{\theta})\sqrt{\frac{\hat{\boldsymbol{F}}^T\hat{\boldsymbol{F}}}{\hat{\sigma}^2}\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}} \tag{25}$$

by noting that $\tilde{e} = \hat{e} + \boldsymbol{F}^T(\tilde{\theta} - \hat{\theta})$. As shown in Appendix III,

$$\frac{1}{r}ln(\frac{r}{q}) \simeq r\frac{3}{4n} \tag{26}$$

and

$$Prob\left(|r| \leq \frac{z_{\alpha/2}}{(1 - \frac{3}{4n})}\right) \simeq 1 - \alpha \tag{27}$$

This result shows that in order to obtain a more accurate coverage probability, the critical value must in fact be increased, with a more substantial correction required for small sample sizes. Confidence intervals formed using this correction will become larger than the approximate intervals, which is expected.

As shown in Appendix III, the correction factor leads to a Student's t approximation with n degrees of freedom for $\frac{(\tilde{\theta}-\hat{\theta})}{s}\sqrt{\hat{\boldsymbol{F}}^T\hat{\boldsymbol{F}}}$. The true distribution is Student's t with $n - 1$ degrees of freedom.

4.1.2 Nonlinear Model

For nonlinear models containing a single parameter θ, it is interesting to relate the q correction to results obtained using the curvature methods introduced by Bates and Watts (1988). In this instance, the expectation mapping can be approximated using a second-order Taylor series expansion about $\hat{\theta}$:

$$\boldsymbol{f}(\theta, \boldsymbol{x}) \approx \hat{\boldsymbol{f}} + (\theta - \hat{\theta})\hat{\boldsymbol{F}} + \frac{(\theta - \hat{\theta})^2}{2}\hat{\boldsymbol{h}} \tag{28}$$

where $\hat{\boldsymbol{f}}$ is the expectation mapping evaluated at the maximum likelihood estimate, and $\hat{\mathbf{h}}$ is the Hessian of this mapping, which in this instance is an $n \times 1$ vector whose i^{th} entry is $\frac{\partial^2 f(\theta, x_i)}{\partial \theta^2}$. Defining $K_T = \hat{\boldsymbol{F}}^T(\hat{\boldsymbol{F}}^T\hat{\boldsymbol{F}})^{-1}\hat{\mathbf{h}}$, it can be established that:

$$q = (\tilde{\theta} - \hat{\theta})\sqrt{\frac{\hat{\boldsymbol{F}}^T\hat{\boldsymbol{F}}}{\hat{\sigma}^2}\left(1 + \frac{K_T}{2}(\tilde{\theta} - \hat{\theta})\right)\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}}$$

and

$$\frac{1}{r}log(\frac{r}{q}) \approx r\frac{1}{4n}\left(3 + ln\left(\frac{w}{a}\right)(1 + \frac{2}{w})\right) \tag{29}$$

where

$$\frac{w}{a} = \frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\tilde{\sigma}_L^2 - \hat{\sigma}^2} \frac{1}{\left(1 + \frac{K_T}{2}(\tilde{\theta} - \hat{\theta})\right)^2}, \qquad w = \frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2} \tag{30}$$

In this equation $\tilde{\sigma}_L^2$ is the maximum likelihood estimate of the variance one would have anticipated had the had the model been linear in the parameters:

$$\tilde{\sigma}_L^2 = \hat{\sigma}^2 + \frac{(\tilde{\theta} - \hat{\theta})^2}{n} \hat{\boldsymbol{F}}^T \hat{\boldsymbol{F}} \tag{31}$$

Equations (29) and (30) indicates that the confidence interval for r will will be asymmetric. This example serves to illustrate that for very simple nonlinear models, the correction factors quickly become complicated functions.

4.2 General Expressions For Coordinate Parameters:

Coordinate parameters are any one of the estimated parameters in the model, including those in the description of the stochastic terms such as the noise variance. For a very broad class of likelihood functions, the correction factor q for a coordinate parameter has the form (Fraser, Reid and Wu (1999); Fraser, Wong and Wu (1999); Brazzale et al. (2007)):

$$q = \frac{\left| \begin{pmatrix} \boldsymbol{\psi}(\hat{\boldsymbol{\theta}}; \hat{\sigma}) - \boldsymbol{\psi}(\tilde{\boldsymbol{\theta}}; \tilde{\sigma}) & \boldsymbol{\psi}_\theta(\tilde{\boldsymbol{\theta}}; \tilde{\sigma})_{red} \end{pmatrix} \right|}{|\boldsymbol{\psi}_\theta(\hat{\boldsymbol{\theta}}; \hat{\sigma})|} \sqrt{\frac{|\hat{\boldsymbol{I}}|}{|\tilde{\boldsymbol{I}}_{red}|}} \tag{32}$$

While this expression looks formidable, the quantities are readily calculated in most instances. Attention to numerical linear algebra is required, and implementation schemes are discussed in Appendix II. Explicit formulae for each of these terms for the case of IID Normal errors and multi-response situations where the stochastic terms are not cross-correlated, are given in Appendix I. Expressions for non-normal error structures and heteroscedastic error structures can be calculated from formulae provided in Fraser, Wong and Wu (1999) and Brazzale et al. (2007). Rather than present the explicit formulaes in the main body, a brief discussion is given:

1. $\boldsymbol{\psi}(\hat{\boldsymbol{\theta}}; \hat{\sigma}), \boldsymbol{\psi}(\tilde{\boldsymbol{\theta}}; \tilde{\sigma})$ are $(p+1) \times 1$ vectors, known as the *canonical parameters*, that are functions of model residuals, and the Jacobian of the expectation function at $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$

2. $\boldsymbol{\psi}_\theta(\hat{\boldsymbol{\theta}}; \hat{\sigma}), \boldsymbol{\psi}_\theta(\tilde{\boldsymbol{\theta}}; \tilde{\sigma})$ are $(p+1) \times (p+1)$ matrices that are functions of the model residuals, and the Jacobian of the function at $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$. $\boldsymbol{\psi}_\theta(\tilde{\boldsymbol{\theta}}; \tilde{\sigma})_{red}$ is a $(p+1) \times p$ matrix obtained by removing the column of $\boldsymbol{\psi}_\theta(\tilde{\boldsymbol{\theta}}; \tilde{\sigma})$ corresponding to the appropriate co-ordinate parameter that is being investigated.

3. $\hat{\boldsymbol{I}}, \tilde{\boldsymbol{I}}$ are $(p+1) \times (p+1)$ matrices, known as observed Information matrices, that are functions of the model residuals, the Jacobian of the expectation mapping at $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ respectively, and the Hessian of the expectation mapping. $\tilde{\boldsymbol{I}}_{red}$ is a $p \times p$ matrix obtained by removing the column and row of $\tilde{\boldsymbol{I}}$ corresponding to the appropriate co-ordinate parameter that is being investigated.

   For the case of an IID normal stochastic element Fraser, Wong and Wu (1999); Brazzale et al. (2007):

$$\tilde{\boldsymbol{I}} = \tilde{\sigma}^2 \begin{pmatrix} \tilde{\boldsymbol{F}}^T \tilde{\boldsymbol{F}} - \sum_{t=1}^{n} \tilde{e}_t \frac{\partial^2 f(\boldsymbol{\theta}, \boldsymbol{x}_t)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} & 2\tilde{\boldsymbol{F}}^T \tilde{\boldsymbol{\varepsilon}} \\ 2\tilde{\boldsymbol{\varepsilon}}^T \tilde{\boldsymbol{F}} & 4n \end{pmatrix} \tag{33}$$

The Hessian at each observation t, is evaluated at the constrained values of the parameters. Recall that the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is the upper left $p \times p$ block of the inverse of $\hat{\boldsymbol{I}}$. When calculating the Information Matrix at the maximum likelihood estimates, the term involving the product of the residuals and the Hessian is most often ignored. Furthermore, the term $\hat{\boldsymbol{F}}^T \tilde{\boldsymbol{\varepsilon}} = \boldsymbol{0}$ since this defines the necessary conditions for minimum of the sums of squares function. Neglecting the Hessian term leads to serious errors in calculation of the constrained information matrix.

Skovgaard (1996) has developed a correction factor that is of order $1/n$ that is similar to Eq(32). Moment-based approximations to the canonical parameters are used. It is still necessary to solve the constrained optimization and to compute the observed Information matrix at the constrained solution.

4.3 General Expression for Differentiable Function of Model Parameters

In many scientific and engineering applications, the ability to determine confidence intervals for functions of parameters is very important. Typical functions of interest might include a prediction of the independent variable at a particular set of input settings, the ratio of this value to the standard error of prediction, critical points, such as maximum, minimum and inflection points in growth models and compartment models. Fraser and co-workers (Fraser, Reid and Wu (1999), Fraser, Wong and Wu (1999)), and Brazzale et al. (2007) provide expressions for the calculation of q for an arbitrary differentiable function of the model parameters. This function is not restricted to functions of $\boldsymbol{\theta}$, but may include components involving the variance-term as well. The expressions are only slightly more involved that those in Eq(32). It has been our experience however that the numerical properties of the method are inferior to those described in the previous section. The reason for this is that the method requires evaluation of $\det \tilde{I}$. This matrix can quickly become ill-conditioned at moderate departures of $\hat{\boldsymbol{\theta}}$ from $\tilde{\boldsymbol{\theta}}$, rendering the results suspect.

Instead, we have found that the results in the previous section can be used directly, if the function of interest is imbedded as an *extra* parameter in the optimization. Bates and Watts (1988) suggested that the profile-t approach could be used to determine likelihood intervals for a function of interest, if the model could be re-parameterized such that expectation function was re-reexpressed in terms of the function. This approach is quite limited in application, difficult to implement and restricted to functions of the model parameters. An alternate approach is to embed the constraint as an extra function as follows:

1. Solve the unconstrained problem to obtain $\{\hat{\boldsymbol{\theta}}, \hat{\sigma}^2\}$

2. Append the model to include a 'pseudo-observation',a 'pseudo-prediction', and an additional parameter as follows:

$$\begin{aligned} y_{N+1} &= 0 \\ f_{N+1}(\boldsymbol{x}, [\boldsymbol{\theta}, \theta_{p+1}]) &= R(g(\boldsymbol{x}, \boldsymbol{\theta}) - \theta_{p+1}) \end{aligned} \tag{34}$$

   In this formulation the extra parameter then corresponds to the constraint. For this method to be successful, the scale parameter, R, must be chosen large enough to ensure that the optimizer forces the right-hand size of Eq(34) to zero. The choice is not difficult. A sensible value is that $R = 100N\hat{\sigma}^2$.

3. The results in the previous section can be used directly, treating the function of the parameters now as a co-ordinate parameter. There are significant computational advantages to using this strategy, and these are discussed in the next section. Note that the intiial value for the 'appended' optimizer uses the starting value $\hat{\theta}_{p+1} = \hat{g} = g(\hat{\boldsymbol{\theta}}, \hat{\sigma})$. Another advantage of this approach is that if the unconstrained optimizer is started with these values, it will will usually converge within one or two iterations with the same parameter values, returning the standard error of the function as a by-product. This is especially convenient when the gradient of the function is evaluated numercially.

4. Bellio et al. (2000) describe the application of higher-order asymptotic corrections proposed by Barnorff-Nielsen and Cox (1986) and Vidoni (1998). These corrections do not use a constrained approach to the likelihood function, are more computationally awkward, and are not considered further in this paper.

5 Computation of the Constrained Optimum

Calculation of the profile-t, profile-r and profile-r*, requires repeated solution of a constrained optimization. There are several ways this can be accomplished. Many nonlinear least squares solver do not have the capability to minimize the objective function subject to an arbitrary constraint. Often however, they do enable one to fix one or more of the parameters at prescribed values. One can exploit this feature to solve the constrained optimization problems for co-ordinate parameters. In section 4 it was shown how to embed an arbitrary constraint as an extra parameter, thus enabling the use optimizers which allow parameters to be fixed.

Instead of solving the nonlinear optimization problem using a nonlinear optimizer, it is possible to solve the constrained optimization by solving the Lagrange equation:

$$\frac{\partial}{\partial(\boldsymbol{\theta},\lambda)^T}\left(S(\boldsymbol{\theta})+\lambda(g(\boldsymbol{x},\boldsymbol{\theta})-c)\right)=\mathbf{0} \tag{35}$$

In this equation, λ is the Lagrange multiplier and c is the constraint value that is to be satisfied. The solution of constrained optimization problems by solving the associated set of nonlinear equations that are associated with Eq(35) has been studied extensively (e.g., Bertsekas (1982)). Applications to profiling have been proposed in Venzon and Moolgavkar (1988), Allgower and Georg (2003), and Chen and Jennrich (2002). The Lagrange equations can be solved with a nonlinear equation solver, or they can be solved using a differential equation solver for the following set of equations:

$$\begin{pmatrix} \ddot{\boldsymbol{S}}(\boldsymbol{\theta}_c)+\lambda_c\ddot{\boldsymbol{g}}(\boldsymbol{x},\boldsymbol{\theta}_c) & \boldsymbol{g(\theta_c)} \\ \boldsymbol{g^T(\theta_c)} & 0 \end{pmatrix} \begin{pmatrix} \dot{\boldsymbol{\theta}}_c \\ \dot{\lambda}_c \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ 1 \end{pmatrix} \tag{36}$$

where $\ddot{\boldsymbol{S}}(\boldsymbol{\theta}_c)$ and $\ddot{g}(\boldsymbol{x},\boldsymbol{\theta})$ denote the second derivative of $\boldsymbol{S}(\boldsymbol{\theta})$ and $g(\boldsymbol{x},\boldsymbol{\theta})$ respectively. These equations are integrated from $c=g(\boldsymbol{x},\hat{\boldsymbol{\theta}})$ to c. Initial conditions are given by $(\hat{\boldsymbol{\theta}},0)$.

Chen and Jennrich (2002) prove that the solution to the Lagrange equations is also obtained by solving the following set of simultaneous differential equations:

$$\begin{pmatrix} -\boldsymbol{W}(\boldsymbol{\theta}_c) & \boldsymbol{g}(\boldsymbol{\theta}_c) \\ \boldsymbol{g}(\boldsymbol{\theta}_c) & 0 \end{pmatrix} \begin{pmatrix} \dot{\boldsymbol{\theta}}_c \\ \dot{\mu}_c \end{pmatrix} = \begin{pmatrix} -\gamma\dot{\boldsymbol{S}}(\boldsymbol{\theta}_c) \\ 1 \end{pmatrix} \tag{37}$$

In this equation $\boldsymbol{W}(\boldsymbol{\theta})$ is a positive-definite matrix, and γ is a 'large' positive constant. There is considerable choice in the selection of these two factors. Ideally, $\boldsymbol{W}(\boldsymbol{\theta})$ is selected as the $\boldsymbol{I}(\boldsymbol{\theta}_c)$. Pseudo-code is provided in Chen and Jennrich (2002). We have found that the selection $\boldsymbol{W}(\boldsymbol{\theta})=\boldsymbol{F}(\boldsymbol{\theta}_c)^T\boldsymbol{F}(\boldsymbol{\theta}_c)$ and $\gamma=0$ is very reliable. When combined with a fixed step size (i.e. an Euler integration method), this corresponds to solving the nonlinear-constrained problems using a sequential-linearzed-constrained least squares problem. Pseudo-code for a numerically sound implementation is given in Golub and van Loan (1996). Eq(37) can be interpreted as an embedded-homotopy solution to the Lagrange equations Garcia and Gould (1980); Allgower and Georg (2003). The inclusion of $-\gamma\dot{\boldsymbol{S}}(\boldsymbol{\theta}_c)$ provides a correction term, enabling larger integration time steps (Chen and Jennrich (2002); Richter and DeCarlo (1983)).

There are important advantages to solving the constrained optimization via the solution of the Lagrange equations. As will be demonstrated in the subsequent section, this approach is extremely rapid. Second, and perhaps more important, is that there are none of the convergence issues that are often associated with the solution to the nonlinear least squares optimization algorithm. This is particularly true for problems where convergence of the algorithm is very sensitive to initial conditions.

Calculation of the higher-order correction (and solution of the Lagrange equations) requires first and second order derivative information. The wide-spread availability of computer-algebra packages provides a convenient path for calculation of these quantities for relatively simple problems. For many problems, the use of these packages is not possible, especially where quantities of interest must be calculated via numerical methods. Numerical derivatives are required. Sophisticated packages are available, that enable sparse-matrix methods to be used, automatic selection of step-size. Some of the most intriguing and accurate approaches use complex arithmetic to compute the derivatives to obtain remarkably accurate numerical derivative (Shampine (2007)). (All of this is transparent however to the user.)

6 Examples

In this section, a number of examples are studied that demonstrate the range of nonlinear behaviour that can be encountered, and illustrate the use and improvement provided by higher-order asymptotic corrections, and to provide a perspective on the numerical computations. All computations were performed in the MATLAB^{TM} environment, using a nonlinear optimizer having a damped Newton strategy that accommodates equality constraints (Nowak and Weimann (1990)). This algorithm is available without license for noncommerical applications. Numerical derivatives were used in the optimizer for all examples, although the algorithm allows the user to provide a function for analytic derivatives. This algorithm provides great flexbility, while having tremendous ease of use. For the higher-order asymptotics, first- and second-order derivatives are required. The first-order derivatives for these corrections were obtained analytically, while the second-order derivatives were obtained numerically from the analytical first-order derivatives using the algorithms described and available in Shampine (2007). In the examples that follow, n refers to the number of observations available, while p refers to the number of parameters in the model.

6.1 Chlorine Degradation: $n = 42, p = 2$

In their text on applied regression analysis, Draper and Smith (1998) provide as an an example a dataset describing the amount of chlorine available in a product used for washing clothes. The amount of chlorine in the product deteriorates over time. In the dataset, the response y is the residual chlorine available, expressed as a fraction of the original chlorine, while the regressor t is the time in weeks that the product has been stored. The following model is estimated from the data:

$$f(\boldsymbol{\theta}, t) = \theta_1 + (0.49 - \theta_1)exp(-\theta_2(t - 8)) \tag{38}$$

The data and predictions from the estimated model,$f(\hat{\boldsymbol{\theta}}, t)$, are plotted in Figure 1.

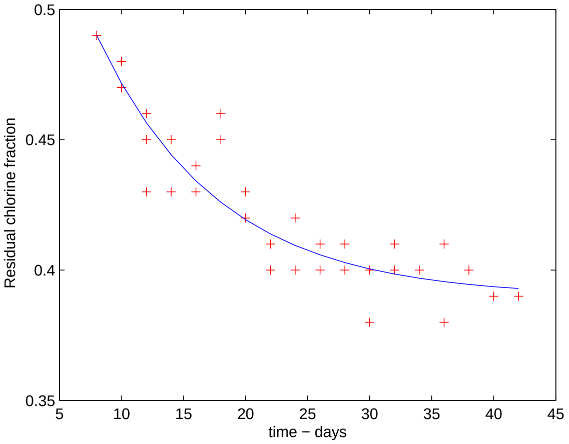

Figure 1: Residual Chlorine and $f(\hat{\boldsymbol{\theta}}, t)$ (data from Draper and Smith Draper and Smith (1998))

This example was studied by Quinn et al. (1999), where it was shown using the profile-t function that the parameters exhibit mild departures from linearity, and no serious errors are introduced by using the linear-approximation-based asymptotic results. The profile-r and profile-r* plots for this example are shown in Figure 2. (Recall that the profile-r* is the profile-r function with the higher-order correction defined in Eq(22)). The reference line (based on linear approximation) is tangent to the profile-r curve at the maximum likelihood estimate, as expected. The reference line will not necessarily be tangent to the profile-r* curve at the maximum likelihood estimate because it is calculated from linearization of the profile-r function. In many cases, discontinuities will be observed in the profile-r* at $\hat{\boldsymbol{\theta}}$, arising from the correction factor q in the r^* expression. In order to provide smooth profile-r* plots at the maximum likelihood parameter estimates, the value of r^* plotted on the graph was obtained by averaging values of r^* at two points adjacent to $\hat{\boldsymbol{\theta}}$. As well, it is entirely possible for the profile-r* curve to be offset from the profile-r curve.

From Figure 2, we would conclude that departures from linearity are small. Both the profile-r and profile-r* curves lie above the reference line, indicating that the sum of squares functions increases less rapidly than would be expected had the model been linear in the parameters. The profile plots for θ_2 (not shown) also exhibit mild departures from linearity, but in this instance like below the reference line.

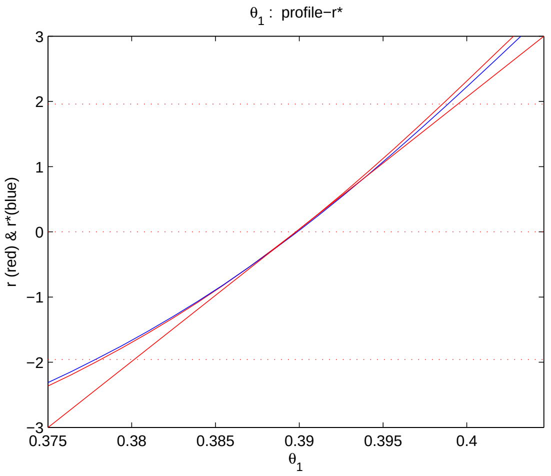

Figure 2: Chlorine Example - Profile r- and Profile r*-plots for θ_1

Likelihood and confidence intervals are summarized in Table 1. The confidence intervals were calculated using as an estimate of the noise variance the MSE from the maximum likelihood estimation, and not the maximum likelihood estimate of σ_ϵ^2. The 95% linearized confidence interval for each parameter is tabulated, and then the upper and lower segments of the likelihood intervals from the profile-r, profile-r*, and profile-t functions are expressed as a percentage of the half-width of the linearized confidence interval. For example, a value of 1.2 for the upper likelihood interval segment indicates that the distance between the upper $\alpha/2$ likelihood segment and the maximum likelihood estimate is 20% larger than than the distance between the $\alpha/2$ linearized confidence interval and the maximum likelihood estimate. A comparison of the likelihood intervals indicates that intervals based on the profile-r, profile-r*, and profile-t functions are all comparable. There is a slight asymmetry in the likelihood intervals for θ_1, as indicated by the percentage differences in the range of $15\% - 18\%$ for the lower segments, versus percentage differences of $8\% - 11\%$ for the upper segments. The linearized confidence intervals are inherently symmetric. There is less asymmetry in the likelihood intervals for θ_2.

Models are often used to determine secondary, or follow-on, quantities of interest. For example, suppose that the washing product is essentially ineffective once the fractional residual chlorine drops below 0.40. The time t_{tc} at which this threshold concentration is reached can be obtained from the model equation, Equation (38):

$$t_{tc} = g(\theta) = \frac{1}{\theta_2} ln \left(\frac{.49 - \theta_1}{.40 - \theta_1} \right) + 8 \tag{39}$$

Such a quantity provides a guideline for storage of the product, and as part of the analysis developing storage guidelines, one would want to develop inference intervals for the maximum storage time. This is a derived quantity of the model parameters, and represents an example of generalized profiling of a function $g(\cdot)$ of the parameters. The likelihood intervals for $g(\theta)$ are very nonlinear as seen in the profile-r and profile-r* plots shown in Figure 3, and in the intervals summarized in Table 1. In particular, regardless of whether the likelihood intervals are based on the profile-t, profile-r and profile-r* functions, the lower and upper segments are respectively in the range of 75% and 250% of the corresponding confidence interval segment. This suggests a dramatic asymmetry in the likelihood intervals about the maximum likelihood parameter estimate. There is a pronounced deviation from linearity visible in the profile plots in Figure 3 especially at the upper end, and this is reflected in the asymmetry in the likelihood intervals relative to the approximate confidence interval. The likelihood interval is defined by the intersection of the dashed lines at $\pm t_{\nu,0.025}$ with the linear reference curve, and the r and $r*$ plots. At the upper end of the interval, the pronounced bending in the curves means that the point of intersection will be at much larger values of t_{tc}, which is reflected in the fact that the upper intervals are 250% of those of the linearized confidence interval.

In order to compute the profile-r* function, the correction factor q was calculated several ways: i) imbedding t_{tc} as a parameter of interest in the parameter estimation problem using the prediction parameter transformation approach of Bates and Watts (1988), ii) including t_{tc} as an extra parameter as dissussed in Section ( ), and iii) using the formula of Fraser, Reid and Wu (1999) directly. The profile-t, profile-r and profile r* (computed using approaches (i) and (ii)) gave nearly identical results, both in terms of the likelihood intervals and in the profile curves. Numerical difficulties were encountered in using approach (iii). The incorporation of the constraint as an extra parameter in method (ii) does not require that the original model be reparameterized, as in method (i), which can be a tremendous advantage in most situations.

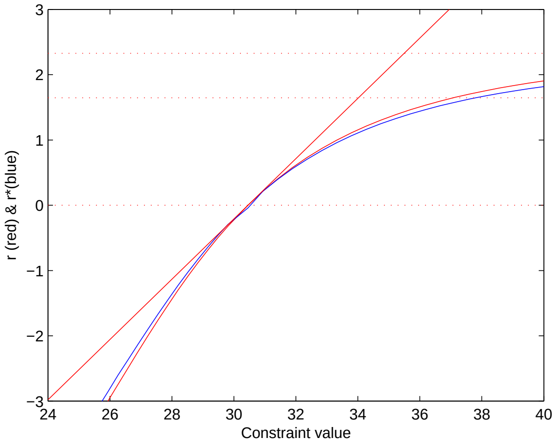

Figure 3: Profile r and $r*$ plots and likelihood intervals for t_{tc}

The importance of reparameteration to reduce nonlinear effects has been advocated in Ratkowsky (1983), Bates and Watts (1981), Watts (1994), Chen and Jennrich (1995), and Chen and Jennrich (2002). Problems exhibiting high nonlinearity are also prone to basic estimation problems, including premature termination of the numerical algorithm). Chen and Jennrich (1995) advocate looking for a monotone-linearizing transformation. They provide constructive algorithms for determining such transformations. If such transformations exist, more accurate confidence intervals (e.g., for the nonlinear function t_{tc} in this example) can be readily obtaining by inverting the transformation. While the form of the constraint function $t_{tc} = g(\boldsymbol{\theta})$ in Eq(39) of this example appears to be amenable to a log-type transformation, it is not.

Box (1971) proposed a measure relating expected bias in an estimate to nonlinearity. Box's measure of expected bias was extended to functions of parameters by Ratkowsky (1983). The expected bias in an estimate of a function of the parameters is defined to be proportional to the trace of the Hessian of the function with respect to the parameters, evaluated at the least squares or maximum likelihood estimates. Chen and Jennrich (1995) defined a relative nonlinearity index γ that is the ratio of the second differential of the function in the direction of the gradient to the norm of the first differential of the function, evaluated at the maximum likelihood or least squares estimate, scaled by the noise standard deviation. Non-zero values of the nonlinearity index γ indicate nonlinearity, while values close to zero indicate a linear problem. The nonlinearity index for t_{tc} is provided in Table 1 and has a value of -0.37, indicating some nonlinearity present.

6.2 Richards Example: $n = 11, p = 4$

The Richards function,Richards (1959), is often used in biological models to describe growth:

$$f(\boldsymbol{\theta}, t) = \theta_4 - \theta_3 ln\left(1 + exp\left(-\frac{\theta_2}{\theta_3} - \frac{\theta_1}{\theta_3}t\right)\right) \tag{40}$$

The expectation function and observed values are plotted in Figure 4 using data from Nelder (1961). The response is proportional to the growth rate of a vegetable product, and t refers to a scaled and translated time, for which negative values are permissible. Units were not provided. Note also that the number of degrees of freedom for this estimation problem is small, with 11 observations and 4 parameters being estimated. Figure 4 also shows the prediction performance of the model with maximum likelihood parameter estimates. From the figure, the estimated model provides a very good prediction of the observed responses.

The impact of parameter nonlinearities in this model was studied in Clarke (1987) using the profile-t approach. All of the likelihood regions show significant departures from their corresponding linearized confidence regions. Focusing on θ_2, Box's percentage estimate of the bias due to curvature is -1.4%, a value that exceeds the 1% magnitude threshold that Ratkowksy notes will usually signal problems in linear-approximation-based inference. The Chen and Jennrich nonlinearity index $\gamma = -0.3981$ and is non-zero, again

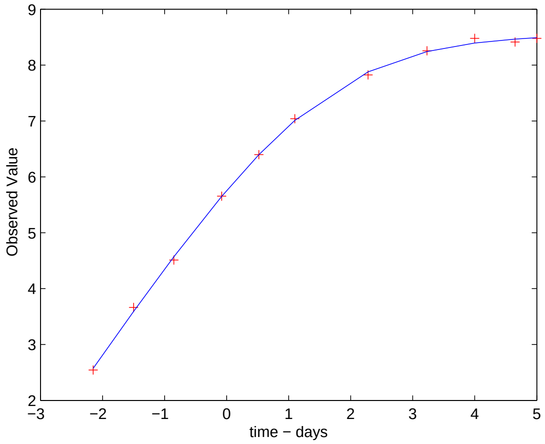

Figure 4: Estimated Richards function using data (shown) from Nelder Nelder (1961)

indicating that there is nonlinearity present in the relationship between θ_2 and the response. Both the Box bias measure and the nonlinearity index suggest that bias and asymmetry may be a problem in inference. Nonlinearity index values are summarized for the remaining parameters in the model in Table 1, from which it can be seen that the largest nonlinearity index is associated with θ_2.

The profile-r and profile-$r*$ plots are shown in Figure 5, together with the linearization reference line. The r and $r*$ curves exhibit curvature and deviation from the reference line, which is more pronounced above the maximum likelihood parameter estimate for the $r*$ curve, which has the higher-order correction. However, below $\hat{\theta}_2$ the $r*$ curve closely matches the linearization curve, whereas the r curve shows deviations from the linearization line both above and below $\hat{\theta}_2$. There is considerable deviation between the r and $r*$ curves, suggesting that the higher-order correction will be important. The linearization-based 95% confidence intervals, along with the likelihood intervals from the profile-t,profile-r and profile-$r*$ methods are shown in Table 1.

In growth models, a common function of interest is the time t_{max} at which the growth rate reaches its maximum value:

$$t_{max} = \frac{-\theta_2 + log(\theta_3)\theta_3}{\theta_1} \tag{41}$$

Confidence and likelihood interval results for t_{max} are are summarized in Table 1, with likelihood interval segments expressed as percentages of the corresponding confidence interval segments. Profile-r and $r*$ profiles are shown in Figure 6. While the results for $r*$ in this figure may appear suspect, they are correct. The correction factor is very influential, and can exhibit a discontinuity at the maximum likelihood estimate. This is not surprising as there are only 7 degrees of freedom. Even if the model were linear, the correction factor

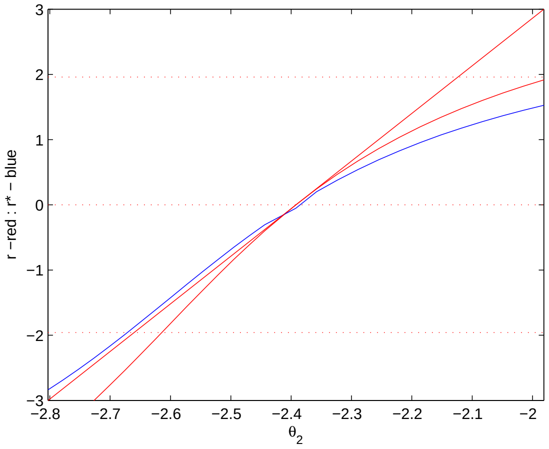

Figure 5: Profile-r (red), profile-$r*$ (blue) functions and likelihood intervals for θ_2

would exhibit similar trends as the correction factor must accomodate the considerable differences between an asymptotical Normal variable and a Student's t distribution with 7 degrees of freedom. In estimating the higher-order correction q that is used to determine the profile-$r*$ plot for θ_2, we have chosen to append an additional parameter. For this example, it is straightforward to reparameterize the model in terms of t_{max} by eliminating θ_2. The likelihood intervals computed using $r*$ closely match those obtained using the profile-t method.

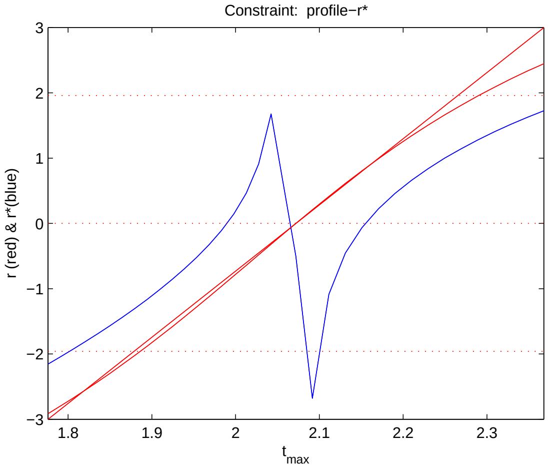

Figure 6: Likelihood intervals for t_{max}

Table 1: Summary of Results: The 95% linearized confidence interval. The upper and lower segments of the likelihood intervals from the profile-r, profile-r*, and profile-t functions are expressed as a percentage of the half-width of the linearized confidence interval

Problem	$\hat{\theta}$	$t - value$	γ	Lineariation		profile-t		profile-r		profile-r*	
				Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Chlorine											
$\hat{\theta}_1$	0.3898	77.2	0.18	0.380	0.400	1.18	0.87	1.15	0.85	1.17	0.89
$\hat{\theta}_2$	0.1017	7.6	-0.10	0.075	0.129	0.91	1.12	0.89	1.08	0.89	1.09
$\hat{t}_{tc}$	30.45	13.7	-0.47	25.98	34.93	0.73	2.57	0.71	2.30	0.74	2.62
Richards											
$\hat{\theta}_1$	1.641	12.2	-0.25	1.427	1.855	0.77	1.41	0.60	0.91	0.75	1.32
$\hat{\theta}_2$	-2.392	5.352	-0.40	-2.798	-1.986	0.70	1.73	0.55	1.05	0.69	1.63
$\hat{\theta}_3$	1.7682	7.336	-0.22	1.058	2.478	0.78	1.38	0.60	0.90	0.75	1.3
$\hat{\theta}_4$	8.5545	6.6481	-0.13	8.4169	8.6921	0.88	1.16	0.65	0.80	0.86	1.10
$\hat{t}_{max}$	2.0719	19	0.3	1.80	2.36	0.90	1.15	0.64	0.75	0.92	1.19

6.3 Viscosity Example: $n = 53, p = 9$

Bates and Watts (1988) include in their text an example of a more complex parameter estimation problem in which viscosity is predicted in terms of excess pressure and temperature. The model equation is:

$$f(\boldsymbol{\theta}, P_t, T_t) = \frac{\theta_1}{\theta_2 + T_t} + \theta_3 P_t + \theta_4 P_t^2 + \theta_5 P_t^3 + (\theta_6 P_t + \theta_7 P_t^3) exp(\frac{-T_t}{\theta_8 + \theta_9 P_t^2}) \tag{42}$$

in which $f(\boldsymbol{\theta}, P_t, T_t)$ represents the logarithm of viscosity, P_t is the excess pressure (bar) and T is the temperature in C. There were 53 observations. Temperature ranges from 0 to 100 C and the pressure ranges from 1 to 10,000 bar. Nonlinearity measures for this problem were discussed in Linssen (1975), and the data are reported in Bates and Watts (1988). A model-building exercise was undertaken to determine the maximum likelihood parameter estimates and to assess the adequacy of the model.

In examining the proposed model, an initial reaction is that the pressure should be logarithmically transformed. This was not fruitful. All of the parameters except θ_9 were assessed as statistically important. The t-values for these parameters all exceeded 9. The residuals are noticeably deficient when any of the parameters, except θ_9 is excluded. The model shows mild indications of heteroscedasticity, but this was ignored in the subsequent analysis. In fitting the model, the pressures and temperatures were scaled by dividing the original values by 100. Convergence of the nonlinear-least squares algorithm was very sensitive to initial values for the parameters.

Linssen (1975) investigated the nonlinearity measures of Beale (1960) for this example. He concluded that the the linearization results might be suspect, and that the nonlinearity effects were strongly directionally dependent. No remedies were proposed. Likelihood intervals for each of the parameters were constructed using the methods described earlier. Figure(7) shows a typical result. One might anticipate that the profile-r and profile-r* results would overlap, since N-p is large. This is not case. The results for the profile-r* are not discernably different from those obtained using the profile-t approach. The chlorine-degradation example illustrated that a model can be effectively linear in the parameters, yet functions can exhibit considerably nonlinearity. First, all of the predictions were profiled using the method of an additional embedded parameter. The predictions were all reasonably linear. Likelihood intervals from the profile-r* were $\sim$ 10% greater than those obtained by neglecting the correction factor.

An important quantity of interest is the viscosity ratio defined as the relative change of viscosity with respect to temperature. Noting that the output variable is given in ln(viscosity), this function is:

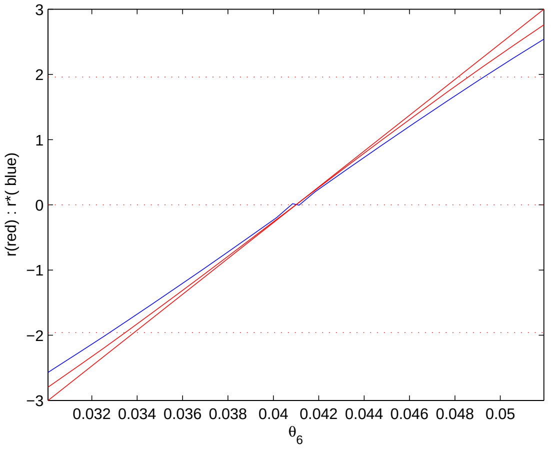

Figure 7: Profile-r (red), profile-$r*$ (blue) functions and likelihood intervals for θ_6

$$\begin{aligned} g(\boldsymbol{\theta}, \boldsymbol{x}) &= \frac{d}{dT} f(\boldsymbol{\theta}, P_t, T_t) \\ &= -\frac{\theta_1}{(\theta_2 + T_t)^2} - \frac{1}{\theta_8 + \theta_9 P_t^2}(\theta_6 P_t + \theta_7 P_t^3) exp(\frac{-T_t}{\theta_8 + \theta_9 P_t^2}) \end{aligned} \tag{43}$$

Another function of interest is the ratio of this quantity to the same function at a reference temperature and pressure.

A sense of the computer time required to produce the profile curves is shown in Table(6.3). The reported time (seconds) is the time to produce the profile curve (40 points) and plot the results for all 53 predictions. A typical desk-top-computing platform was used. The time is reported for the profile-t using the nonlinear optimizer and an Euler integration with Eq(37) and for the profile-r* using the nonlinear optimizer and integrating the equations with the MATLAB^{TM} stiff-equation solver - ode23s. In both cases $\boldsymbol{W}(\boldsymbol{\theta}) = \boldsymbol{F}(\boldsymbol{\theta}_c)^T \boldsymbol{F}(\boldsymbol{\theta}_c)$ and $\gamma = 0$. Interchange of these algorithms did not materially affect the computation time. The nonlinear optimizer used numerical derivatives. Analytic first-order derivatives were provided for the profile-r*. Second-order derivatives were calculated numerically using the first-order derivatives. These results show that i) the use of the homotopy algorithm is much faster than use of the nonlinear optimizer, and ii) that the inclusion of the correction factor requires additional computing time. While the computation time differences may not be important, we have found that the homotopy algorithm is robust and very reliable.

Table 2: Computation Time for Example 3 - time in seconds

	profile-t		profile-r*	
	optimizer	Eq(37)	optimizer	Eq(37)
model predictions	133	15	363	111
viscosity ratio	133	13	444	177

7 Discussion and Conclusions

The use of higher-order methods to adjust the nominal significance level in in likelihood-based inference approaches has been investigated for the case of nonlinear regression models with homoscedastic error variance. These techniques were discussed in the context of inference for scalar-valued functions of parameters, including predictions. The computations can be organized in a straightforward and efficient and computationally stable manner. A new approach for computing these corrections as constrained optimization problems was proposed, building on the previous work of Fraser, Brazzale, and co-workers. Alternative expressions for the higher-order asymptotic corrections were developed that provide a more transparent comparison between corrections for coordinate parameters and those for scalar-valued functions of the coordinate parameters. A comprehensive discussion of numerically efficient and stable computational approaches was provided. The use of homotopy methods for solving the constrained optimization problem was seen to be very effective. The techniques were demonstrated using a sequence of examples of increasing complexity that affords comparison of the efficacy of the various approximate inference approaches. It was observed that the likelihood-intervals provided by the higher-order methods closely matched those obtained using the profile-t approach of Bates and Watts. A theoretical justification was provided for this in the case of a single parameter. The extension to a vector of parameters is currently being investigated.

We also recognize that the higher order corrections developed apply to a much broader scope of problems beyond those for which the profile-t is appropriate. We have not discussed handling of heteroscedastic errors - which was done in Brazzale et al. (2007) et al. and Bellio et al. (2000). As noted by Bellio and Brazzale (2003), the combination of computer algebra and organized modules (routines) will probably be required for widespread adoption of these methods.

We have discussed in this paper the use of resampling techniques such as Monte-Carlo method. Are Monte-Carlo methods more attractive? They are reasonably straightforward to apply for homoscedastic errors, however they are more problematic for models with a heteroscedastic error structure. For smaller, less complicated models, Monte-Carlo methods provide a more convenient approach in the homoscedastic case, however as the models become more complex and where larger data sets are used, the computational load increases. Moreover, when the models are nonlinear, the optimization may bog down in generating the estimates for each Monte-Carlo run, The comparative advantages of Monte-Carlo methods versus higher order asymptotics for large-scale complex models is an unresolved issue. Bellio et al. (2000) comment that on examples that they studied, the corrected intervals provided substantial improvement to those based on the asymptotic normal results. They also noted that the r* (based on the Skovgaard approximation), was conservative in that the coverage probabilities for quantities of interest were between 96 and 97% for a nominal 95% test level. We have performed extensive Monte-Carlo simulations and would reach conclusions that in all cases, the coverage probabilities of the intervals computed using the corrected quantities were much closer to the nominal probabilities than those obtained from linear approximations.

8 Appendix I: Higher Order Correction Factors

The class of models for which higher-order corrections is very large. In this Appendix, we will simplify the expressions in Fraser, Wong and Wu (1999) and Brazzale et al. (2007) for the case of an additive model and error structure of the form:

$$y_t = f(\boldsymbol{x_t}, \boldsymbol{\theta}) + w(\boldsymbol{x_t}, \boldsymbol{\theta}, \sigma)e_t \qquad (t = 1, 2, ...n) \tag{A1-1}$$

where the stochastic term is iid Normal. The vector of parameters, $\boldsymbol{\theta}$ now includes additional terms used to parameterize the dependence of the variance of the stochastic term on explanatory variables. The inclusion of a multi-response data set with m dependent variables $Y_{i,t}, i = 1..m, t = 1..n$ whose stochastic elemenents are contemporaneously independent is readily incorporated by 'stacking' the observations and modelling the variance of each group of responses as $\gamma_i^2\sigma^2, i = 2, \ldots, m$. $\gamma_i, i = 2, \ldots, m$ are additional parameters to be included in the vector of parameters to be estimated. For a model description of the form given by Eq. (A1-1), the sums of squares function in Eq(4) maximization of the likelihood leads to minimization of the following sums of squares function:

$$S(\boldsymbol{\theta}) = \sum_{t=1}^{n} \left(\frac{y_t - f(\boldsymbol{x_t}, \boldsymbol{\theta})}{w_t} \right)^2 \tag{A1-2}$$

where $w_t = w(\boldsymbol{x_t}, \boldsymbol{\theta}, \sigma)$.

For the case of homoscedastic errors, the correction factors can be derived expressions given in Brazzale et al. (2007) and Fraser, Wong and Wu (1999).

Define:

$$\tilde{\boldsymbol{F}} = \frac{\partial \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\boldsymbol{T}}} \mid_{\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}}$$

and

$$\boldsymbol{\psi}(\hat{\boldsymbol{\theta}};\hat{\sigma}) = -\frac{1}{\hat{\sigma}}\begin{pmatrix} \mathbf{0}_p \\ \\ n \end{pmatrix} \tag{A1-3}$$

$$\boldsymbol{\psi}(\tilde{\boldsymbol{\theta}};\tilde{\sigma}) = -\frac{1}{\tilde{\sigma}}\begin{pmatrix} \hat{\boldsymbol{F}}^T \\ \\ \boldsymbol{\varepsilon}^{\hat{\boldsymbol{T}}} \end{pmatrix}\tilde{\boldsymbol{\varepsilon}} = -\frac{1}{\tilde{\sigma}}\begin{pmatrix} \hat{\boldsymbol{F}}^T\tilde{\boldsymbol{\varepsilon}} \\ \\ \hat{\boldsymbol{\varepsilon}}^T\tilde{\boldsymbol{\varepsilon}} \end{pmatrix} \tag{A1-4}$$

$$\boldsymbol{\psi}_\theta(\hat{\boldsymbol{\theta}};\hat{\sigma}) = \frac{1}{\hat{\sigma}^2}\begin{pmatrix} \hat{\boldsymbol{F}}^T \\ \\ \boldsymbol{\varepsilon}^{\hat{\boldsymbol{T}}} \end{pmatrix}\begin{pmatrix} \hat{\boldsymbol{F}} & 2\boldsymbol{\varepsilon}^{\hat{\boldsymbol{T}}} \end{pmatrix} = \frac{1}{\hat{\sigma}^2}\begin{pmatrix} \hat{\boldsymbol{F}}^T\hat{\boldsymbol{F}} & \mathbf{0} \\ \\ \mathbf{0} & 2\boldsymbol{\varepsilon}^{\hat{\boldsymbol{T}}^T}\boldsymbol{\varepsilon}^{\hat{\boldsymbol{T}}} \end{pmatrix} \tag{A1-5}$$

$$\boldsymbol{\psi}_\theta(\tilde{\boldsymbol{\theta}};\tilde{\sigma}) = \frac{1}{\tilde{\sigma}^2}\begin{pmatrix} \hat{\boldsymbol{F}}^T \\ \\ \boldsymbol{\varepsilon}^{\hat{\boldsymbol{T}}} \end{pmatrix}\begin{pmatrix} \tilde{\boldsymbol{F}} & 2\boldsymbol{\varepsilon}^{\tilde{\boldsymbol{T}}} \end{pmatrix} = \frac{1}{\tilde{\sigma}^2}\begin{pmatrix} \hat{\boldsymbol{F}}^T\tilde{\boldsymbol{F}} & 2\hat{\boldsymbol{F}}^T\tilde{\varepsilon} \\ \boldsymbol{\varepsilon}^{\hat{\boldsymbol{T}}^T}\tilde{\boldsymbol{F}} & 2\boldsymbol{\varepsilon}^{\hat{\boldsymbol{T}}^T}\boldsymbol{\varepsilon}^{\tilde{\boldsymbol{T}}} \end{pmatrix} \tag{A1-6}$$

The terms with 'tilde' are evaluated using the parameters at the constrained optima. $\boldsymbol{\psi}(\tilde{\boldsymbol{\theta}};\tilde{\sigma})$ and $\boldsymbol{\psi}_\theta(\tilde{\boldsymbol{\theta}};\tilde{\sigma})$ involve terms evaluated at both the unconstrained and constrained optima. In these expressions, $\hat{\boldsymbol{\varepsilon}}$, denotes the standardized residual from the unconstrained optimization and $\boldsymbol{\varepsilon}^{\tilde{T}}$ denotes the standardized residual using the constrained parameter values. In these expressions, the residual variance is maximum likelihood estimate and is obtained by dividing the residual sums of squares by n and not $n-p$.

Often in multiresponse estimation, the 'relative' variances will be known, and the $\gamma_i, i = 2,\ldots,m$ are not estimated. In this case the above formulae can be used directly by an appropriate pre-scaling of the observed values and expectation function.

9 Appendix II: Numerical Linear Algebra for Higher Order Correction Factors

The general expression for the correction factor is given in Eq(32). This correction factor can be written as:

$$q = \pm w_i\frac{|\boldsymbol{\psi}_\theta(\tilde{\boldsymbol{\theta}};\tilde{\sigma})|}{|\boldsymbol{\psi}_\theta(\hat{\boldsymbol{\theta}};\hat{\sigma})|}\sqrt{\frac{|\hat{\boldsymbol{I}}|}{|\tilde{\boldsymbol{I}}_{red}|}} \tag{A2-1}$$

where w_i is the i^{th} element of the vector $\boldsymbol{w}$ that is the solution to the following set of linear equations

$$\boldsymbol{\psi}_\theta(\tilde{\boldsymbol{\theta}};\tilde{\sigma})\boldsymbol{w} = \boldsymbol{\psi}(\hat{\boldsymbol{\theta}};\hat{\sigma}) - \boldsymbol{\psi}(\tilde{\boldsymbol{\theta}};\tilde{\sigma}) \tag{A2-2}$$

This is a new expression for the correction factor than previously published, and is obtained by recalling Cramer's rule for the solution of linear equations and recalling that the determinant of a matrix obtained by interchanging any two columns is determined to a scaling factor of ±1. Consider the following linear equation, where $\boldsymbol{J}$ is an $p \times (p+1)$ matrix:

$$\boldsymbol{\psi}_\theta(\tilde{\boldsymbol{\theta}};\tilde{\sigma})\boldsymbol{J} = \begin{pmatrix} \boldsymbol{\psi}(\hat{\boldsymbol{\theta}};\hat{\sigma}) - \boldsymbol{\psi}(\tilde{\boldsymbol{\theta}};\tilde{\sigma}) & \boldsymbol{\psi}_\theta(\hat{\boldsymbol{\theta}};\hat{\sigma}) \end{pmatrix} \tag{A2-3}$$

If $\boldsymbol{J}$ is partitioned into the $p \times 1$ vector and a $p \times p$ matrix $\boldsymbol{J} = (\boldsymbol{J_A}|\boldsymbol{J_B})$, then:

$$\begin{aligned} w_i &= (\boldsymbol{J_A})_i \\ \frac{|\boldsymbol{\psi}_\theta(\tilde{\boldsymbol{\theta}};\tilde{\sigma})|}{|\boldsymbol{\psi}_\theta(\hat{\boldsymbol{\theta}};\hat{\sigma})|} &= |\boldsymbol{J_B}| \end{aligned} \tag{A2-4}$$

q and r have the same sign (Brazzale et al. (2007), Fraser, Wong and Wu (1999)) and so the indeterminacy with respect to sign is resolved by taking the absolute value of the quantity $\frac{r}{q}$. While one can effectively solve the problem in this formulation, there are numerical advantages to exploiting the natural partitioned structure of the Information matrices to obtain the determinants. The matrices $\boldsymbol{\psi}(\tilde{\boldsymbol{\theta}};\tilde{\sigma}), \boldsymbol{\psi}_\theta(\hat{\boldsymbol{\theta}};\hat{\sigma}), \boldsymbol{\psi}_\theta(\tilde{\boldsymbol{\theta}};\tilde{\sigma})$ are highly structured. Computational efficiencies and numerical advantage are obtained by using qr decompositions on the main block terms.

10 Appendix III: Limiting Distribution for Higher Order Correction Factor

To further explore the the correction factor for the case of a linear model with one parameter, define:

$$t_0^2 = (\tilde{\theta} - \hat{\theta})^2 \frac{\hat{\boldsymbol{F}}^T \hat{\boldsymbol{F}}}{\hat{\sigma}^2} \tag{A3-1}$$

Then straightforward to show that:

$$r^2 = nlog(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}) = nlog(1 + \frac{t_0^2}{n}) \tag{A3-2}$$

and

$$\frac{1}{r^2} log(\frac{r^2}{q^2}) = \frac{1}{nlog(1+\frac{t_0^2}{n})} log[\frac{(1+\frac{t_0^2}{n})^2}{\frac{t_0^2}{n}} log(1 + \frac{t_0^2}{n})] \tag{A3-3}$$

Expanding the right hand side series in powers of n one obtains:

$$\frac{1}{r^2}log(\frac{r^2}{q^2}) = \frac{1}{n}(\frac{3}{2} - \frac{1}{24}\left(\frac{t_0^2}{n}\right) + \frac{1}{48}\left(\frac{t_0^2}{n}\right)^2 - \frac{13}{960}\left(\frac{t_0^2}{n}\right)^3 + O\left(\frac{t_0^2}{n}\right)^4) \tag{A3-4}$$

Retaining only the term involving n^{-1}:

$$\frac{1}{r}log(\frac{r}{q}) \simeq r\frac{3}{4n} \tag{A3-5}$$

Therefore:

$$Prob\{r(1-\frac{3}{4n}) \leq z^{\alpha/2}\} \simeq 1-\alpha/2 \tag{A3-6}$$

or

$$Prob\{r \leq \frac{z^{\alpha/2}}{(1-\frac{3}{4n})}\} \simeq 1-\alpha/2 \tag{A3-7}$$

The effect of the correction factor is to inflate the confidence interval, a result that was expected. We know the true distribution of $t' = \sqrt{\frac{n-1}{n}}t_0$ is Student-t with (n-1) degrees of freedom. Since the logarithm is a monotonic function, it follows that:

$$\begin{aligned} Prob\{r^2 \leq [\frac{z^{\alpha/2}}{(1-\frac{3}{4n})}]^2\} &\simeq 1-\alpha \\ Prob\{|t'| \leq \sqrt{(n-1)(-1+exp([\frac{z^{\alpha/2}}{\sqrt{n}(1-\frac{3}{4n})}]^2)\}} &\simeq 1-\alpha \end{aligned} \tag{A3-8}$$

Expanding the term in brackets, one obtains:

$$\begin{aligned} \sqrt{(n-1)(-1+exp([\frac{z^{\alpha/2}}{\sqrt{n}(1-\frac{3}{4n})}]^2)\}} &= z^{\alpha/2}[1+\frac{1+[z^{\alpha/2}]^2}{4n} + \frac{4+28[z^{\alpha/2}]^2+5[z^{\alpha/2}]^4}{96n^2}...] \\ &= \simeq z^{\alpha/2}[1+\frac{1+[z^{\alpha/2}]^2}{4n}] \end{aligned} \tag{A3-9}$$

A series similar that on on the right-hand side was studied by Fisher (1926):

$$t_n^{\alpha/2} = z^{\alpha/2}[1+\frac{1+[z^{\alpha/2}]^2}{4n} + \frac{3+16[z^{\alpha/2}]^2+5[z^{\alpha/2}]^4}{96n^2}...] \tag{A3-10}$$

The quality of the first order correction of Fisher (1926), that is, retaining only terms to order $1/n$

has been studied by Scott and Smith (1970). They demonstrate that critical values, $t_n^{\alpha/2}$, obtained with this approximation are very close to the exact values for commonly encountered significance values. The correction factor induced by higher-order method produces an first order confidence interval for a t-variate with n degrees of freedom, instead of $n-1$ degrees of freedom.

References

Allgower E.L and K. Georg. *Introduction to Numerical Continuation Methods.* SIAM, New York, 2003.

Barndorff-Nielson, O. Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika*, 73:307-322, 1986.

Barndorff-Nielsen O.E. and D.R. Cox. Edgeworth and saddlepoint approximations with statistical applications. *J. Royal Stat. Soc: Series B*, 41:279-312, 1984.

Barndorff-Nielsen O.E. and D.R. Cox. Prediction and asymptotics. *Bernoulli*, 2:319-340, 1996.

Bates D.M. and D.G. Watts. Parameter transformations for improved approximate confidence regions in nonlinear least squares. *Applied Statistics*, 9:1152–1167, 1981.

Bates D.M. and D.G. Watts. *Nonlinear Regression Analysis and Its Applications.* John Wiley & Sons, New York, 1988.

Beale E.M.L. Confidence regions in non-linear estimation. *J. R. Statist. Soc. B*, 22:41–76, 1960.

Bellio R. and A.R. Brazzale. Higher-order asymptotics unleashed: Software design for nonlinear heteroscedastic models. *J. CGS*, 12:682-697, 2003.

Bellio,R. and J.E. Jensen and P. Seiden. Applications of likelihood asymptotics for nonlinear regression in herbicide bioassays. *Biometrics*, 56:1204-1212, 2000.

Bertsekas D.P. *Constrained Optimization and Lagrange Mutliplier Methods.* Academic Press, New York, 1982.

Box M.J. Bias in nonlinear estimation. *J. R. Statist. Soc. B*, 32:171-201, 1971.

Brazzale A.R., A.C. Davison, and N. Reid. *Applied Asymptotics: Case Studies in Small-Sample Statistics.* Cambridge University Press, Cambridge, 2007.

Chen J.S. *Confidence Intervals for Parametric Functions in Nonlinear Regression.* PhD thesis, University of California, Los Angeles, CA, 1991.

Chen J.S. and R.I. Jennrich. Diagnostics for linearization confidence intervals in nonlinear regression. *JASA*, 90:1068–1074, 1995.

Chen J.S. and R.I. Jennrich. The signed root deviance profile and confidence intervals in maximum likelihood analysis. *JASA*, 91:993–998, 1996.

Chen J.S. and R.I. Jennrich. Simple accurate approximations of likelihood profiles. *J. Computational and Graphical Statistics*, 11:714–732, 2002.

Clarke G.P.Y. Approximate confidence limits for a parameter function in nonlinear regression. *JASA*, 82:221–230, 1987.

Cook R.D. and S. Weisberg. Confidence curves in nonlinear regression. *JASA*, 85:544–551, 1990.

Cox D.R. and D.V. Hinkley. *Theoretical Statistics.* Chapman and Hall, London, England, 1974.

Donaldson J.R. and R.B. Schnabel. Computational experience with confidence regions and confidence intervals for nonlinear least squares. *Technometrics*, 29:67–82, 1987.

Draper N.R. and H. Smith. *Applied Regression Analysis* Wiley, New York, 1998.

Fisher, R.A. Expansion of 'Student's' integral in powers of n^{-1}. *Metron*, 5:109–112, 1926.

Fraser D.A.S. and J.Wong and J. Wu. Regression analysis, nonlinear or nonormal: Simple and accurate p values from likelihood analysis. *JASA*, 94:1286–1295, 1999.

Fraser D.A.S. and N. Reid, and J. Wu. A simple general formula for tail probabilities for frequentist and bayesian inference. *Biometrika*, 86:249–264, 1999.

Gallant A.R. *Nonlinear Statistical Models*. Wiley, New York, NY, 1987.

Garcia C.B. and F.J. Gould. Relations between several path following algorithms and local and global newton methods. *SIAMReview*, 22(3):263–274, 1980.

Golub G.H. and C.F. VanLoan. *Matrix Computations*. John Hopkins University Press, Baltimore, 3 edition, 1996.

Hamilton D.C. and D.G Watts, and D.M Bates. Accounting for intrinsic nonlinearity in non-linear inference regions. *Annals of Statistics*, 10:386–393, 1982.

Lam R.L.H. and D.G. Watts. Profile summaries for arima time series model parameters. *J. Time Ser. Anal.*, 12:225–235, 1991.

Linssen H.N. Nonlinearity measures: a case study. *Statistica Neerlandica*, 29:93–99, 1975.

Nelder J.A. The fitting of a generalization of the logistic curve. *Biometrics*, 17:89–110, 1961.

Nowak U. and L. Weimann. A family of newton codes for systems of highly nonlinear equations - algorithm, implementation, application. Technical Report TR 90-10, Zuse Institute, Berlin, December 1990.

Quinn S.L. and D.W. Bacon and T.J. Harris. Assessing the precision of model predictions and other functions of model parameters. *Can. J. Chem. Eng.*, 77:723–737, 1999.

Quinn S.L. and D.W. Bacon and T.J. Harris. Notes on likelihood intervals and profiling. *Commun. Statist.-Theory Meth.*, 29:108–130, 2000.

Quinn S.L.and T.J. Harris and D.W. Bacon. Measuring uncertainty in control-relevant statistics. *J. Proc. Cont.*, 15:675–690, 2005.

Ratkowsky D.A. *Nonlinear Regression Modeling*. Marcel Dekker, New York, NY, 1983.

Reid N. Likelihood and higher-order approximations to tail areas: a review and annotated bibliography. *Canadian J. Statistics*, 24:141-166, 1996.

Richards J.F. A flexible growth function for empirical use. *J. Experimental Botany*, 10:290–300, 1959.

Richter S.L. and R.A. DeCarlo. Continuation methods: Theory and applications. 30:347–352, 1983.

Scott, A. and M.F. Smith. A Note on Moran's Approximation to Student's t. *Biometrika*,3:681–682

Seber G.A.F. and C.J. Wild. *Nonlinear Regression*. John Wiley and Sons, New York, NY, 1989.

Severini T.A. and J.G. Staniswalis. Quasi-likelihood estimation in semiparametric models. *JASA*, 89:501–511, 1994.

Shampine L.F. Accurate numerical derivatives in matlab. *TOMS*, 33(4):26–43, 2007.

Skovgaard J.M. An explicit-large deviation approximation to one-parameter tests. *Bernouli*, 2:145–165, 1996.

Skovgaard J.M. Likelihood asymptotics. *SJS*, 28:3-32, 2001

Strawderman R.L. Higher-order asymptotic approximation: Laplace, Saddlepoint and Relaed Methods. *JASA*, 95:1358-1364, 2000.

Venzon D.J. and S.H. Moolgavkar. A method for computing profile-likelihood-based confidence intervals. *Applied Statistics*, 37:87–94, 1988.

Vidoni, P. A note on modified estimative prediction limits and distributions. *Biometrika*, 85:949-953, 1998.

Watts D.G. Estimating parameters in nonlinear rate equations. *Can. J. Chem. Eng.*, 72:701–710, 1994.