Measuring and Predicting Uncertainty in Control-Relevant Statistics

by

Shannon Leigh Quinn

A thesis submitted to the Department of
Chemical Engineering in conformity with the requirements for
the degree of Doctor of Philosophy

Queen's University
Kingston, Ontario, Canada
September, 1999

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

0-612-42969-5

Canada

QUEEN'S UNIVERSITY AT KINGSTON

SCHOOL OF GRADUATE STUDIES AND RESEARCH

PERMISSION OF CO-AUTHORS

I/we, the undersigned, hereby grant permission to microfilm any material designated as being co-authored by me/us in the thesis copyrighted to the person named below:

Shannon L. Quinn	
Name of copyrighted author	**Signature of copyrighted author**
Name(s) of co-author(s)	**Signatures of co-author(s)**
Thomas J. Harris	
David W. Bacon	

DATE: July 19, 1999

Abstract

The broad topic of statistical inference is examined, with a view to reliably estimating the uncertainty in functions of parameters of importance in process control. The method of profiling (Bates and Watts, 1988; Chen, 19991; Lam and Watts, 1991, Chen and Jennrich, 1996) is examined in detail, and is generalized so that it might be used in the context of control-relevant statistics.

The thesis is a compilation of manuscripts, all of which examine and extend the idea of profiling in a chemical engineering context. Nonetheless, their subjects are quite diverse. The papers touch on issues of inference, measuring nonlinearity and design of experiments. The models considered range from nonlinear regression models to discrete dynamic transfer function models. Contributions have been made to the disciplines of applied statistics, chemical engineering and control theory.

The first two manuscripts are focused on consolidating the theory of generalized profiling. The first paper is a tutorial on the use of profiling to estimate reliable likelihood intervals for functions of parameters, and illustrates the method using examples involving nonlinear regression models. In the second paper, the equivalence of two approaches to generalized profiling is shown, and cases for which profiling fails are identified. An alternative to profiling is suggested for one of these cases.

Dynamic models are introduced in the third manuscript. In this paper, a method called *expected profiling* is developed. Expected profiling is a tool for predicting the uncertainty in functions of parameters of time series models, which does not require a set of data. It may be used to examine how quickly the uncertainty in a function of parameters is expected to decrease as the length of the data set increases, and to estimate how much data is required for asymptotic properties to apply effectively.

In the fourth paper, two new measures of nonlinearity are introduced, one for time series models and one for nonlinear regression models. Measures of nonlinearity are intimately related to measuring uncertainty in functions of parameters since it is the nonlinearity of the response surface that complicates the inference problem. The new measure of nonlinearity for autoregressive moving average (ARMA) time series models is based on the fact that the proximity of the vector of parameter values to a stability/invertibility boundary largely influences the degree of nonlinearity of the inference problem. Both new measures of nonlinearity are "quick and easy" methods to predict when iterative inference methods, such as generealized profiling, are required.

The final paper is devoted to exploring the use of generalized profiling in the context of transfer function models for the purpose of measuring uncertainty in control-relevant statistics. The lessons learned in the earlier manuscripts are used to discuss the inference problems which provided the initial motivation for this thesis.

Acknowledgments

To my mother, who has always supported me, and who supported me in this work despite believing it had something to do with canning peas;
to my father, who has come to know every pothole between Kingston and Ottawa, and who would seemingly drive any distance to pick me up;
to my brother, whose kindness and wit inspire me;
to Dante, who loves me;
to Jim, who always had time to help me with anything;
to Tom, for his genius, his encouragement and his energy;
to Dr. Bacon, who has been a friend, a colleague and a mentor;
and to all of the students in the Department who have made my time at Queen's memorable,

thank you.

Co-Authorship

The research for the manuscripts, which constitute Chapters 3 to 7 of this thesis, was carried out under the supervision of Drs. T.J. Harris and D.W. Bacon. Chapter 3 was submitted to the *Canadian Journal of Chemical Engineering*, and has been accepted for publication. Dr. Bacon, Dr. Harris and I appear as co-authors of this paper. Chapter 4 has been submitted for publication and is currently under review. The authorship is the same as for the first paper. The material in Chapters 5, 6 and 7 has been presented, in part, at national and international conferences.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The use of nonlinear models is widespread in chemical engineering, and indeed, in the natural and applied sciences in general. Modeling can serve a wide range of purposes, one of which is the prediction of response values at specific levels of the independent (or explanatory) variables. If these predictions are to be used appropriately for decision making, then reliable measures of the uncertainty in the predictions must be known. Model predictions are but one example of functions of parameters of interest in engineering. In this thesis, numerical procedures for the estimation and prediction of reliable likelihood intervals for functions of parameters of nonlinear models are developed. The work illustrates the application of these techniques to chemical engineering problems, with an emphasis on statistics of interest in process control.

Three main topics are considered: estimating likelihood intervals for functions of parameters, predicting the uncertainty in functions of parameters of time series models, and measuring the extent to which inference problems will display nonlinearity. Although, to some extent, each of these topics is distinct, they each stem from, or lead to, the other. At their core, they are all tied to the issue of measuring uncertainty in functions of parameters of nonlinear models.

To make appropriate use of fitted models, whether they be deterministic or stochas-

tic, mechanistic or empirical, the uncertainty in the statistics must be known. For models which are linear in the parameters, inference results for the parameters and the predictions are exact and analytic. For models which are nonlinear in the parameters, the inference results are complex functions of: the distribution of the random errors associated with the measured responses, the structure of the model, the parameterization of the model, and the design of the experiment (Bates and Watts, 1988). Exact inference results are unavailable, except for special cases (see, for example: Williams, 1962; Halperin, 1963; Roy, 1992). Typically, approximate inference results have been based on linear approximations to the nonlinear model. This linear approximation approach is attractive in that it provides quick analytic results. However, Bates and Watts (1980), Donaldson and Schnabel (1987), Ratkowsky (1990), and others, have shown that the linear approximations are often poor, and that either the intrinsic or parameter effects nonlinearity (or both) are often severe enough to render the linearization results unreliable and possibly misleading.

A more reliable approach to estimating inference results for the statistics of a nonlinear model is profiling. Profiling, as developed by Bates and Watts (1988), is a graphical likelihood ratio approach to finding likelihood intervals for the parameters of a nonlinear model. Subsequent work by Cook and Weisberg (1990), Lam and Watts (1991), Chen (1991) and Chen and Jennrich (1996) generalized the profiling algorithm to handle the problem of finding likelihood intervals for functions of parameters of fitted models for which the distribution of the random errors is known.

In this thesis, the generalized profiling algorithm will be examined in detail so as to elucidate and consolidate the theory underlying the method, and thereby to assess its power and limitations. Generalized profiling can be used to solve extremely varied inference problems because many statistics of practical interest can be conceived of as being functions of parameters. For example, the expected responses of a model (i.e., the predictions from a model) are functions of the parameters in the model. The

usefulness of profiling in the context of transfer-function models will be investigated, with the focus being on statistics used in process control. Although profiling is a powerful approach to inference problems, the algorithm may fail in special cases. These limitations will be investigated, and alternative approaches will be suggested.

While generalized profiling is a reliable means by which to estimate uncertainty in functions of parameters once a set of data has been used to fit a model, in some cases it would be useful and advisable to predict the uncertainty of estimated parameters and functions of parameters prior to data collection or experimentation. In the case of ARMA models, the model itself defines how the process is expected to evolve over time, and it is possible to calculate expected values for some of the properties of the model *a priori* to any data collection (i.e., in the absence of a realization of the process). An algorithm for computing "expected likelihood intervals" for functions of parameters of ARMA models was developed. Several useful ways of plotting the information obtained from the expected profiling methodology are proposed, and are intended to emphasize the utility of expected profiling in designing experiments. Especially when one has control over how much data can be collected, expected profiling can be an important means for deciding how much data will be "enough". Even for cases where there is little control over data collection, expected profiling may be used as a means of deciding which approaches to computing inference results are appropriate, and for making qualitative judgments about the sources of any observed nonlinearities.

Generalized profiling is a more reliable approach to estimating inference results for functions of parameters of proposed models than the linearization approximation approach, but it is also more computationally intensive. When the nonlinearity of the inference problem is low, it is appropriate to use the linearization approach to estimate uncertainty. Measures of nonlinearity are developed to indicate when it is appropriate to save the computational burden of profiling by using the simpler

linearization approach. Measures of nonlinearity, then, are intimately tied to the issues of inference in the context of nonlinear models. In this work, two empirical measures of nonlinearity are developed - one for use with ARMA models, and one for use with nonlinear regression models. Both of the measures are relatively easy to calculate and are reliable indicators of nonlinearity.

The thesis is organized in "manuscript" format. The main body of the thesis comprises a collection of five journal-ready papers, the topics of which are those introduced above. Naturally, there exists some duplication of information from one paper to another because each paper is meant to be able to stand alone. Still, each paper is distinct and emphasizes different methods, different applications, and/or different approaches. Each paper makes its own unique contribution.

The nomenclature used in each chapter is specific to that chapter and is defined in the text and in a "Nomenclature" section at the end the chapter. The notation for each paper was chosen to be consistent with that used by the discipline to which it was aimed. Because the conventions used in statistics, control and chemical engineering differ, the notation is not consistent from chapter to chapter.

The first paper is a general introduction to generalized profiling in the context of nonlinear regression models. It is a tutorial-type paper with an emphasis on the application of the methodology to chemical engineering problems.

The second paper is of a more academic nature. The main result is a proof of the equivalence of two different approaches to generalized profiling proposed in the literature. Together with the first paper, cases are identified for which the profiling methodology fails, and an alternative methodology is proposed.

The third paper introduces the concept of expected profiling. The methodology is developed and illustrated. The application of the algorithm is discussed, and the use of the delta parameterization is used as an example of expected profiling of functions of parameters.

The fourth paper provides the development of two new measures of nonlinearity - one for ARMA models, and one for nonlinear regression models. The advantages and uses of these methods are illustrated.

The fifth paper brings the focus back to application and chemical engineering problems. The paper examines in detail the use of generalized profiling in the context of discrete transfer function models, and illustrates the methodology for several functions of parameters of interest in process control.

A chapter providing a general literature review precedes the papers and is intended to provide a review of the topics fundamental to this work. Each paper provides a literature review specific to its topic. Therefore, Chapter 2 entitled Background and Literature Review, remains quite general. The focus is on the geometrical aspects of the work. Appendix A provides a summary of the computational algorithms used in the thesis, and other important details about computational issues. The thesis ends with a summary of the main conclusions of each paper, and a discussion of directions for future work.

Chapter 2

Background and Literature Review

Each manuscript contained in this thesis includes a "Background and Literature Review" section. It is not the intention to reproduce all of this information here. The focus of this chapter will be exploring ideas common to all of the manuscripts so as to motivate the individual topics and highlight the commonalities that unite them in a single thesis.

The two themes common to all five manuscripts are nonlinearity and inference. In many engineering applications, a nonlinear model is one which is nonlinear is the states, input variables, or regressor variables. However, in the statistical literature, a nonlinear model is one which is nonlinear in the parameters. It is parametric nonlinearity which influences inference results. In this work, nonlinear will always be used in the statistical sense.

A model is said to be nonlinear if the vector of derivatives of the model with respect to its parameters is a function of one or more of these parameters. In the work that follows we consider several classes of models including nonlinear regression models, autoregressive moving average (ARMA) time series models and transfer function

models. These, and other single response models, may be written generally as:

$$y = f(\boldsymbol{x}, \boldsymbol{\theta}) + \epsilon \tag{2.1}$$

where the function $f(\boldsymbol{x}, \boldsymbol{\theta})$ is the expected value of the response variable y at specified levels of m independent variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$ and specified values of p parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$. ϵ is an additive random error term associated with y. At this point assume only that the error is independently and identically distributed (iid) without assuming the form of the distribution. We only consider models which are differentiable to order 2 at least. Model (2.1) is nonlinear if $\frac{\partial f(\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = h(\boldsymbol{\theta})$, where $h(\boldsymbol{\theta})$ is a vector of derivatives, one or more of which is a function of $\boldsymbol{\theta}$.

Time series models and transfer function models may be encompassed within this general framework by allowing $\boldsymbol{x}$ to include past values of y and ϵ. Consider for example, the transfer function model

$$A(q^{-1})y_t = B(q^{-1})u_{t-d} + C(q^{-1})\epsilon_t \tag{2.2}$$

where q^{-1} is a backshift operator such that

$$q^{-1}y_t = y_{t-1} \tag{2.3}$$

and $A(q^{-1})$, $B(q^{-1})$ and $C(q^{-1})$ are polynomials having the form:

$$A(q^{-1}) = 1 + a_1 q^{-1} + \ldots + a_{na} q^{-na} \tag{2.4}$$

$$B(q^{-1}) = b_0 + b_1 q^{-1} + \ldots + b_{nb} q^{-nb} \tag{2.5}$$

$$C(q^{-1}) = 1 + c_1 q^{-1} + \ldots + c_{nc} q^{-nc} \tag{2.6}$$

and d is the delay between a change in the manipulated variable and its affect on y_t. This model may be rewritten as:

$$\begin{aligned} y_t &= ([1 - A(q^{-1})]y_t + B(q^{-1})u_{t-d} + [C(q^{-1}) - 1]\epsilon_t) + \epsilon_t & (2.7) \\ &= \left(\frac{B(q^{-1})}{C(q^{-1})}u_{t-d} + \left[1 - \frac{A(q^{-1})}{C(q^{-1})}\right] y_t\right) + \epsilon_t & (2.8) \end{aligned}$$

so that for this model

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \left(\frac{B(q^{-1})}{C(q^{-1})}u_{t-d} + \left[1 - \frac{A(q^{-1})}{C(q^{-1})}\right] y_t\right) \tag{2.9}$$

and

$$\boldsymbol{x} = (y_{t-1}, \ldots, y_{t-na}, u_{t-d}, \ldots, u_{t-d-nb}) \tag{2.10}$$

ARMA time series models are a special case of (2.2) where $B(q^{-1}) = 0$. For an ARMA(p,q) model,

$$\phi(q^{-1})y_t = \theta(q^{-1})\epsilon_t \tag{2.11}$$

the expressions for the derivatives of ϵ_t with respect to the parameters are (Ravishanker, 1994):

$$\begin{aligned} \frac{\partial \epsilon_t}{\partial \phi_k} &= -\frac{1}{\phi(q^{-1})}\epsilon_{t-k} = -\frac{1}{\theta(q^{-1})}y_{t-k} & (2.12) \\ \frac{\partial \epsilon_t}{\partial \theta_l} &= \frac{1}{\theta(q^{-1})}\epsilon_{t-l} = \frac{\phi(q^{-1})}{\theta^2(q^{-1})}y_{t-l} & (2.13) \end{aligned}$$

where in (2.12) and (2.12) ϕ_k represents the k^{th} parameter of the polynomial $\phi(q^{-1})$,

and θ_l represents the l^{th} parameter of the polynomial $\theta(q^{-1})$. Note that the derivatives are functions of $\boldsymbol{\theta}$; therefore, general ARMA models are nonlinear in the parameters. By extension, the transfer function model given in (2.2) is also nonlinear in the parameters, as are more general forms of transfer function models. The exception is the ARX model having the form $A(q^{-1})y_t = B(q^{-1})u_{t-d} + \epsilon_t$, and any model which is contained within this framework. These models can be shown to be linear in the parameters. Refer to Chapter 7 and the book by Söderström and Stoica (1989) for details.

Donaldson and Schnabel (1987) found that inference results for the parameters of regression models were significantly affected by the nonlinearity of the model. Others, including Bates and Watts (1988) and Ross (1990) have documented similar evidence.

It is common in engineering to base all inference results on linear approximations to the model. This is an attractive option because inference results for linear models have analytic expressions, making linear approximation inference computationally simple. However, the experience of Donaldson and Schnabel suggests that this is an unreliable approach to inference. Watts (1994) showed that the parameters used routinely in chemical engineering may show severe nonlinearity, but that reliable likelihood intervals for those parameters can be obtained by profiling. Profiling is an iterative approach to inference (Bates and Watts, 1988). Bates and Watts (1991) showed how to use profile t plots in model building and model discrimination. Data sets from chemical engineering were used to illustrate the value of profiling in this context. The work by Watts (1994) and Bates and Watts (1991) motivated the research that follows, and served as a starting point for exploring other inference problems of importance in chemical engineering.

2.1 Functions of Parameters

Watts (1994) considered the problem of estimating the uncertainty in the parameters of steady-state rate equations. His work suggested a need for a reliable approach to estimating the uncertainty in functions of parameters of nonlinear models, including time series models and discrete transfer function models. Often in chemical engineering, it is not the parameters of an estimated model which are of primary interest, but rather, statistics derived therefrom. For example, model predictions, measures of controller performance, and the cross-over frequency, are three functions of parameters which are important statistics used for decision-making by chemical engineers. The examples considered in this work are taken from the area of process control, although examples abound in other areas of chemical engineering, and indeed, throughout engineering in general.

2.2 Uncertainty, Likelihood and Confidence

If decisions are to be based on the estimated value of a function of parameters $g(\boldsymbol{\theta})$, then not only is the point estimate of $g(\boldsymbol{\theta})$ important, but so too is a measure of the uncertainty of that estimate. We seek an algorithm by which to obtain a likelihood interval for $g(\boldsymbol{\theta})$. A nominal $1-\alpha$ likelihood interval for $g(\boldsymbol{\theta})$ is the set of all values of $g(\boldsymbol{\theta})$ which are plausible in light of the available data. From standard asymptotic arguments (Cox and Hinkley, 1974)

$$LI\left(g(\boldsymbol{\theta})\right) = \left\{ g(\boldsymbol{\theta}) : -2\ln\left(\frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})}\right) \leq \chi^2_\alpha(1) \right\} \tag{2.14}$$

where $LI(g(\boldsymbol{\theta}))$ is the likelihood interval for $g(\boldsymbol{\theta})$, $L(\boldsymbol{\theta})$ is the likelihood function for $\boldsymbol{\theta}$, $\chi^2_\alpha(1)$ is the upper α quantile for the χ^2 distribution with 1 degree of freedom, $\hat{\boldsymbol{\theta}}$ is the vector of maximum likelihood estimates of the parameters, and $\boldsymbol{\theta}$ is any allowable

vector of parameter values (Chen and Jennrich, 1996). Note that the likelihood ratio statistic follows asymptotically a χ^2 distribution (Cordeiro et al., 1994; Edwards, 1972), except in special cases for which it is exact. The expression in (2.14) states that the likelihood interval for $g(\boldsymbol{\theta})$ includes all values of $g(\boldsymbol{\theta})$ for which $-2\ln\left(\frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})}\right)$ is less than or equal to a critical value of the chi-squared distribution with one degree of freedom. In order to also account for uncertainty in the value of the variance of the random error σ^2, we compute likelihood intervals based on

$$LI\left(g(\boldsymbol{\theta})\right) = \left\{ g(\boldsymbol{\theta}) : -2\ln\left(\frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})}\right) \leq F(1, n-p; \alpha) \right\} \tag{2.15}$$

where $F(1, n-p; \alpha)$ is the upper α quantile for the F distribution with 1 and $n-p$ degrees of freedom (Cook and Weisberg, 1990), and n is the number of observations of the response variable used to compute $L(\boldsymbol{\theta})$. In other words, a $1-\alpha$ likelihood interval for $g(\boldsymbol{\theta})$ is the set of all $g(\boldsymbol{\theta})$ for which

$$-t(n-p, \alpha/2) \leq \sqrt{-2\ln\left(\frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})}\right)} \leq t(n-p, \alpha/2) \tag{2.16}$$

where $t(n-p, \alpha/2)$ is the upper $\alpha/2$ quantile for the student t distribution with $n-p$ degrees of freedom (Bates and Watts, 1988). The likelihood interval is similar to the confidence interval, the fiducial interval (Fisher, 1935) and the Bayesian interval (Jeffreys, 1948) in that all attempt to identify the plausible values of a random variable $g(\boldsymbol{\theta})$. In this work, we restrict our attention to likelihood intervals and approximate confidence intervals. The term "confidence interval" has a precise statistical definition rooted in the frequency theory of probability (Kendall and Stuart, 1967). To construct a $(1-\alpha)100\%$ confidence interval for $g(\boldsymbol{\theta})$ we choose upper and lower confidence limits,

$\tilde{g}_{low}$ and $\tilde{g}_{up}$, respectively, such that

$$P(\tilde{g}_{low} < g(\boldsymbol{\theta}) < \tilde{g}_{up}) = 1 - \alpha \tag{2.17}$$

where the probability is based on the sampling distribution of some function of a random sample of values from a population. A confidence interval is one way to report the results of a test of hypothesis. The Principle of Maximum Likelihood states that, when confronted with a choice of hypotheses, choose the one which maximizes the likelihood (Kendall and Stuart, 1967). This is to say, choose the hypothesis which gives the greatest probability to the set of observations. This is not the same as choosing the hypothesis with the greatest probability (Kendall and Stuart, 1967). Bates and Watts (1988) discussed the difference between the frequency theory results and the likelihood results from a geometrical perspective. The frequency approach to constructing inference regions focuses on the possible values of $f(\boldsymbol{x}, \boldsymbol{\theta})$ and the angle that the residual vector $\boldsymbol{e} = \boldsymbol{y} - f(\boldsymbol{x}, \boldsymbol{\theta})$ makes with the surface traced out by $f(\boldsymbol{x}, \boldsymbol{\theta})$ in n-dimensional space (see Section 2.3.1 for a discussion of this surface). The likelihood approach is centered around the vector of observed values $\boldsymbol{y}$ and its position relative to the expectation surface traced out by $f(\boldsymbol{x}, \boldsymbol{\theta})$. The limits of the likelihood interval are determined by comparing the length of the shortest distance from $\boldsymbol{y}$ to $f(\boldsymbol{x}, \boldsymbol{\theta})$ and the distance from $\boldsymbol{y}$ to all other points $f(\boldsymbol{x}, \boldsymbol{\theta})$. In other words, when constructing confidence intervals, we consider the intersection of a sphere centered at $f(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$ with the expectation surface, but when constructing likelihood intervals, we consider the intersection of a sphere centered at $\boldsymbol{y}$ with the expectation surface. Also, note that we base likelihood intervals on the fact that $-2\ln(LR)$ is asymptotically distributed as a χ^2 random variable. It is possible to construct cases for which it is inadvisable to assume this distribution and for which a likelihood interval approach to constructing inference intervals will give misleading or spurious results. Freund

and Walpole (1987, p. 407) work through such an example.

For linear models, the likelihood interval and the confidence interval are equal and exact. For nonlinear models, the two intervals may not be equal. Because the term "confidence interval" is sometimes used quite loosely in the literature, we will use the more appropriate terms "likelihood interval" and "likelihood region" to describe inference intervals and regions based on profiling.

Although the coverage probability (the value of the confidence coefficient $1-\alpha$) for a likelihood interval is not exact (in most cases), the true coverage probabilities are generally much closer to the nominal values than are those for inference intervals based on a linear approximation to the model (Cook and Weisberg, 1990). The coverage probabilities for likelihood intervals are independent of parameter-effects nonlinearity (see Section 2.3.1). Also, these intervals are exact in shape since they are based on the function $f(\boldsymbol{x}, \boldsymbol{\theta})$. Whereas in the linear approximation interval the function is evaluated at only one point $\hat{\boldsymbol{\theta}}$, and all inferences are based on this and the slope of $f(\boldsymbol{x}, \boldsymbol{\theta})$ at this point, likelihood intervals are formed by computing the value of the likelihood function over a range of values of $\boldsymbol{\theta}$. As a consequence, the function $f(\boldsymbol{x}, \boldsymbol{\theta})$ is evaluated over a range of values of $\boldsymbol{\theta}$ and the likelihood intervals are thereby based on the true nonlinear surface traced out by $f(\boldsymbol{x}, \boldsymbol{\theta})$.

2.3 The Basics of Profiling

The work reported in this thesis has been carried out using the method of generalized profiling. In its most general form, profiling involves the solution of the following constrained optimization problem:

Maximize:

$$L(\boldsymbol{\theta}) \tag{2.18}$$

subject to:

$$g(\boldsymbol{\theta}) = c$$

This optimization problem is solved for a series of values of c greater than and less than the maximum likelihood estimate of $g(\boldsymbol{\theta})$. By the invariance of the likelihood function under reparameterization (Kendall and Stuart, 1967), the maximum likelihood estimate of $g(\boldsymbol{\theta})$ is $g(\hat{\boldsymbol{\theta}})$. Let the location of the solution to the constrained optimization problem in (2.18) be $\tilde{\boldsymbol{\theta}}$, and define

$$\tau(g(\tilde{\boldsymbol{\theta}})) = sign(g(\tilde{\boldsymbol{\theta}}) - g(\hat{\boldsymbol{\theta}}))\sqrt{-2\ln\left(\frac{L(\tilde{\boldsymbol{\theta}})}{L(\hat{\boldsymbol{\theta}})}\right)} \tag{2.19}$$

Then the profile t plot for $g(\boldsymbol{\theta})$ can be constructed as a plot of $\tau(g(\tilde{\boldsymbol{\theta}}))$ versus $g(\tilde{\boldsymbol{\theta}}))$ for a range of values of c. 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 inference results would be. A reference line, $\delta(g(\tilde{\boldsymbol{\theta}}))$ versus $g(\tilde{\boldsymbol{\theta}})$, is typically included on profile t plots, where

$$\delta(g(\tilde{\boldsymbol{\theta}})) = \frac{g(\tilde{\boldsymbol{\theta}}) - g(\hat{\boldsymbol{\theta}})}{se(g(\hat{\boldsymbol{\theta}}))} \tag{2.20}$$

This reference line may be used to obtain the linearization confidence intervals for $g(\boldsymbol{\theta})$, and to judge its relative curvature (Chen, 1991). By (2.16), the likelihood interval includes all values of $g(\tilde{\boldsymbol{\theta}})$ for which

$$-t(n-p, \alpha/2) \leq \tau(g(\tilde{\boldsymbol{\theta}})) \leq t(n-p, \alpha/2) \tag{2.21}$$

To obtain a $(1-\alpha)$ likelihood interval for $g(\boldsymbol{\theta})$ from the profile t plot, one simply finds the values of $g(\boldsymbol{\theta})$ on the profile t curve at which $\tau = \pm t(n-p; \alpha/2)$.

One purpose of this thesis is to provide a comprehensive analysis of generalized profiling. The idea of extending profiling to compute inference results for functions of model parameters has been suggested by Clarke (1987), Bates and Watts (1988), Cook and Weisberg (1990), Ross (1990), Chen (1991), and Chen and Jennrich (1996). Appropriate theory and computational methods have been developed by Clarke (1987), Chen (1991), and Chen and Jennrich (1996). Two apparently different approaches have been proposed: one is based on a reparameterization of the model, and the other involves the constrained optimization formulation given in (2.18). Both methods, although equivalent (see Chapter 4), are conceptually and computationally different. The intent of this work is to elucidate and consolidate the theory underlying both methods, to assess the relative advantages and limitations of each method, and to indicate the usefulness of the methods for several chemical engineering problems.

2.3.1 The Expectation Surface and Nonlinearity

Although the linearization approach provides a computationally simple means of finding inference intervals, it can be misleading (Donaldson and Schnabel, 1987; Ratkowsky, 1983; Bates and Watts, 1980). To appreciate why and how the nonlinearity of a model impacts the inference results, it is necessary to understand the geometry of nonlinear regression.

Bates and Watts (1980, 1988) have provided a detailed analysis of the geometry of the nonlinear regression problem. The discussion which follows relies heavily on these works. Consider a regression problem in which a model having p parameters is fitted to a data set of length n. In an n-dimensional space, where each of the axes represents one of the observations in the data set, the vector $\boldsymbol{y}$, representing the set of observed response values, is a single point in the n-dimensional space. As the p parameters are allowed to vary over their allowable values, the expectation function $f(\boldsymbol{X}, \boldsymbol{\theta})$ traces out a p-dimensional surface in the n-dimensional space, where $\boldsymbol{X}$ is

an $(m \times n)$ matrix of the set of n levels of each of m regressor variables at which the observations were taken. This surface is called the expectation surface or the solution surface. The maximum likelihood estimate of $\boldsymbol{\theta}$ is the value of $\boldsymbol{\theta}$ which defines the point on the expectation surface closest to the point representing the observed values.

The projection of $\boldsymbol{y}$ onto the surface is $\hat{\boldsymbol{y}}$, and the value of $\boldsymbol{\theta}$ which defines this point consists of the least squares estimates of the parameters $\hat{\boldsymbol{\theta}}$. Inferences about $\boldsymbol{\theta}$ are also straightforward, and uncertainty can be represented in the n-dimensional space as a sphere centered at $\boldsymbol{y}$. The size of the sphere is a function of the distance from $\boldsymbol{y}$ to the point $f(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$, the number of degrees of freedom of the estimate of the variance of the random error in the data, and the degree of confidence $(1-\alpha)$. The intersection of the sphere and the surface defines a likelihood region. These ideas are illustrated in Figure 2.1 for the case of a linear model and three observations. For linear models, the expectation surface is planar, making projections and intersections readily calculable. The geometry of regression can also be viewed within a p-dimensional parameter

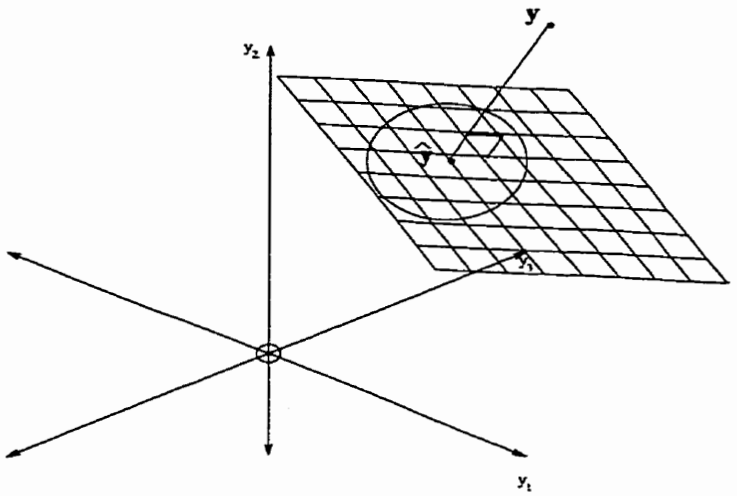

Figure 2.1: The Geometry of Linear Regression (Bates and Watts, 1988).

space. When the model is linear, the circular inference region on the solution surface maps to an elliptical region in a rectangular coordinate region for the parameters. The geometry in these cases is neat in that all of the inference regions are well defined

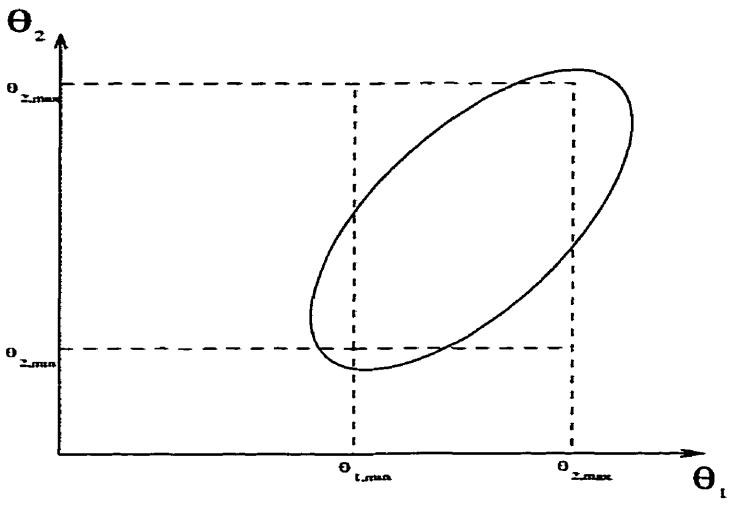

Figure 2.2: A Typical Elliptic Joint Confidence Region for Two Parameters of a Linear Model.

conic structures. This allows for exact, analytic results for both the point estimates of parameters and the inferences.

For nonlinear models, the solution surface is not planar, and the mapping of the likelihood inference region from the surface to the rectangular parameter space is nonlinear. For irregular solution surfaces and highly nonlinear mappings, the likelihood regions for the parameters may be asymmetric and even discontinuous. Therein lies the difficulty in the nonlinear estimation problem. These issues are illustrated by the following example.

Consider an experiment for which only two measurements of y were made, and the measurements are to be used to estimate the single parameter in the following model

$$y = 9500(1 - e^{-\theta_1 t}) \tag{2.22}$$

The data are given in Table 2.1. This estimation problem is displayed graphically in Figure 2.3. The vector of observations $\boldsymbol{y}^T = (7950, 8050)$ is shown on the figure as x, and the location of the vector of best model predictions (in the least squares sense) is the point on the expectation surface (the curved line) closest to x. This point is labeled

Table 2.1: Data for Illustrative Example 1

Time t	Pressure y
2	7905
4	8050

o. The value of the parameter which defines this point is $\hat{\theta}_1 = 0.7835$, the maximum likelihood estimate of the parameter. A likelihood interval for θ_1 includes all values of θ_1 which define points on the expectation curve within a specified distance of $\boldsymbol{y}$. This distance is a function of the level of confidence α, the variance of the random error in the data, and the number of degrees of freedom associated with the estimate of that variance. On Figure 2.3 the limits of the likelihood interval for θ_1 are defined by the points of intersection of the circle centered at x and the expectation curve. Notice that the upper limit extends beyond the limit of the expectation curve. For this model, the expectation surface has a finite upper bound because as θ_1 approaches infinity, the value of the expectation function approaches an upper limit of 9500. The fact that the circle, which serves to define the likelihood interval, extends beyond the upper limit of the expectation curve implies that the upper bound of the likelihood interval for θ_1 is infinity. The + signs on the expectation curve represent equally spaced values of the parameter θ_1. Notice that the ticks are not equally spaced on the expectation curve. This is a result of the nonlinearity of the mapping from the expectation surface to the parameter space and vice versa. Bates and Watts (1980) called this type of nonlinearity parameter-effects curvature. Intrinsic curvature refers to the nonlinearity of the expectation surface itself. For nonlinear modeling problems, both the parameter-effects curvature and the intrinsic curvature complicate estimation and inference.

For the current example, the linearization confidence interval would be misleading. The linearization results would be based on a line tangent to the expectation curve

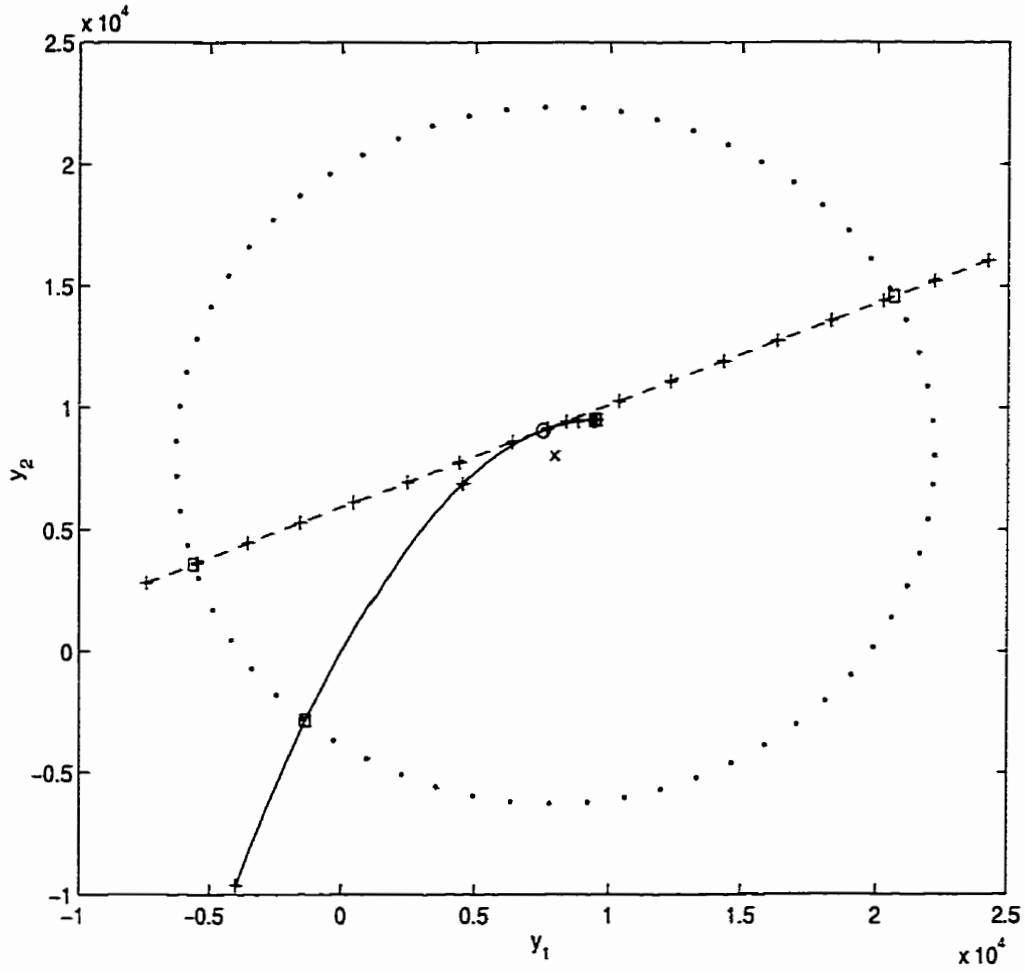

Figure 2.3: A Geometrical Representation of the Estimation Problem in Example 1. – expectation curve; - - linear approximation line; o $\hat{y}$; x y; □ points defining the limits of the likelihood interval.

at $\hat{y}$. This is shown in Figure 2.3 as the dashed straight line. Note that the linear approximation to the model also linearizes the mapping from the parameter space to the space of the observations so that equally spaced values of θ_1 in the parameter space appear as equally spaced ticks on the linear approximation to the expectation line. Based on the linear approximation, the confidence interval for θ_1 has a finite upper bound of 4.105, and the linear approximation interval gives no hint of the true amount of uncertainty in the estimate of θ_1.

The linear inference results can be interpreted from a linear algebra perspective in terms of eigenvectors and eigenvalues. First, identify the values of all parameters

when one parameter θ_q is at one of the limits of its confidence interval as determined using the linearization approach. The boundary of the joint confidence region for the parameters is defined by:

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{V}^T \boldsymbol{V} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = p s^2 F_\alpha(p, n-p) \tag{2.23}$$

This expression defines an ellipse whose principal axes are defined by the eigenvectors of $\boldsymbol{V}^T\boldsymbol{V}$. From the geometry, the locations of the limits of the linear confidence intervals for each parameter lie along one of these eigenvectors. Note that when individual confidence intervals, as opposed to joint regions, are of interest, the appropriate scaling of the contours is $s^2 F_\alpha(1, n-p)$ and not $ps^2 F_\alpha(p, n-p)$ (Donaldson and Schnabel, 1987) . Also, for an individual parameter θ_q, the limit of its $(1-\alpha)$ confidence interval can be obtained from

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = \lambda (\boldsymbol{V}^T \boldsymbol{V})^{-1} \boldsymbol{a} \tag{2.24}$$

where $\boldsymbol{a} = \frac{\partial \theta_q}{\partial \boldsymbol{\theta}} = [0, \ldots, 0, 1, 0, \ldots, 0]^T$ and the 1 is in row q, and λ is a constant which defines the confidence level. For a linear function of the parameters, $g(\boldsymbol{\theta}) = \boldsymbol{a}^T\boldsymbol{\theta}$, (2.24) can still be used to define the locations of the limits of a confidence interval for $g(\boldsymbol{\theta})$.

Substituting (2.24) into (2.23):

$$\begin{aligned} \lambda \boldsymbol{a}^T (\boldsymbol{V}^T \boldsymbol{V})^{-1} \boldsymbol{V}^T \boldsymbol{V} \lambda (\boldsymbol{V}^T \boldsymbol{V})^{-1} \boldsymbol{a} &= s^2 F_\alpha(1, n-p) \\ \lambda^2 \boldsymbol{a}^T (\boldsymbol{V}^T \boldsymbol{V})^{-1} \boldsymbol{a} &= s^2 F_\alpha(1, n-p) \end{aligned} \tag{2.25}$$

Therefore,

$$\lambda = \sqrt{\frac{s^2 F_\alpha(1, n-p)}{\boldsymbol{a}^T (\boldsymbol{V}^T \boldsymbol{V})^{-1} \boldsymbol{a}}} \tag{2.26}$$

The geometrical discussion has so far focused on parameter estimation and inference about the parameters. However, the primary focus of this thesis is inference about functions of parameters. Now consider a function of parameters $g(\boldsymbol{\theta})$ in the geometrical framework.

Since $g(\boldsymbol{\theta})$ is often not a function of the measured data (although it could be), consider the geometry of the inference problem in the space of the parameters $\boldsymbol{\theta}$. In maximum likelihood estimation of $\boldsymbol{\theta}$, we solve the unconstrained optimization problem: maximize $L(\boldsymbol{\theta})$. In finding a likelihood interval for $g(\boldsymbol{\theta})$ by the method of generalized profiling, we solve the constrained optimization problem given in (2.18). In generalized profiling, we then compute $\tau(g(\tilde{\boldsymbol{\theta}}))$. If $|\tau| < t(n-p;\alpha/2)$, we increase c, and revisit the optimization problem in (2.18). This iterative process is continued until $\tau = t(n-p;\alpha/2)$. At this point, $g(\boldsymbol{\theta}) = c = \hat{g} + \Delta$ is one limit of the likelihood interval. The other limit is found by considering a series of negative increments in c beginning at $c = g(\hat{\boldsymbol{\theta}})$. Figure 2.4 illustrates the procedure geometrically. Contours

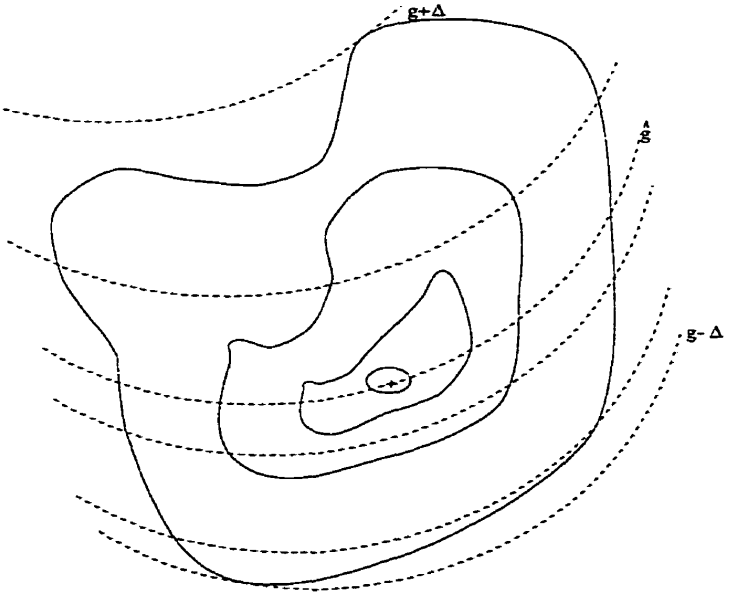

Figure 2.4: A Geometrical Representation of Profiling $g(\boldsymbol{\theta})$.

of $L(\boldsymbol{\theta})$ and $g(\boldsymbol{\theta})$ are shown on the figure. The maximum likelihood estimate of g is equal to the level of the contour of g which passes through the point of maximum

likelihood of $\boldsymbol{\theta}$, i.e., $g(\hat{\boldsymbol{\theta}})$. The first step in profiling $g(\boldsymbol{\theta})$ is to move away from the $g(\hat{\boldsymbol{\theta}})$ contour onto the contour of g having the value $\hat{g}+\Delta$. A search for the maximum of $L(\boldsymbol{\theta})$ *along this contour* of g is then carried out. The upper limit of the likelihood interval for g is g_{max}, and is the largest value of $g(\boldsymbol{\theta})$ that still lies on or within the likelihood region for $\boldsymbol{\theta}$. The likelihood region for $\boldsymbol{\theta}$ in this case is defined by

$$LR(\boldsymbol{\theta}) = \left\{ \boldsymbol{\theta} : -2 \ln \left(\frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})} \right) \leq F(1, n-p; \alpha) \right\} \tag{2.27}$$

When $f(\boldsymbol{x}, \boldsymbol{\theta})$ and $g(\boldsymbol{\theta})$ are both linear in $\boldsymbol{\theta}$, the likelihood region for $\boldsymbol{\theta}$ is an ellipse and the contours of $g(\boldsymbol{\theta})$ are parallel straight lines (see Figure 2.5(a)). The limits of the likelihood interval for $g(\boldsymbol{\theta})$ are those values of the contours of $g(\boldsymbol{\theta})$ which are tangent to the appropriate likelihood region for $\boldsymbol{\theta}$.

The confidence interval for any linear function of the parameters $g(\boldsymbol{\theta}) = \boldsymbol{a}^T\boldsymbol{\theta}$ is:

$$g(\hat{\boldsymbol{\theta}}) \pm st(n-p; \alpha/2) \sqrt{\boldsymbol{a}^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{a}} \tag{2.28}$$

where $\boldsymbol{a}$ is a $p \times 1$ vector of constants. Note that the expected value of a prediction at $\boldsymbol{x}_k$ from a linear model is a linear function of the parameters $\boldsymbol{a}^T\boldsymbol{\theta} = \boldsymbol{x}_k\boldsymbol{\theta}$.

Inference about $g(\boldsymbol{\theta})$ becomes significantly more complicated when $g(\boldsymbol{\theta})$ or $f(\boldsymbol{x}, \boldsymbol{\theta})$, or both, are nonlinear. Some examples of such situations are illustrated in Figure 2.5 (b), (c) and (d). Figure 2.5 (d) illustrates the case for which $g(\boldsymbol{\theta})$ is not monotonic, and for which $g(\boldsymbol{\theta})$ reaches an unconstrained optimum at $\hat{\boldsymbol{\theta}}$. This is an interesting example because for such cases the profiling algorithm fails. This topic is considered in detail in Chapters 3 and 4.

It is obvious that nonlinearity complicates the matter of finding inference results for $g(\boldsymbol{\theta})$. However, it should be emphasized that even if a model is highly nonlinear in the parameters, it does not necessarily follow that all functions of these parameters will also behave nonlinearly. For example, it has been held (Clarke, 1987) that

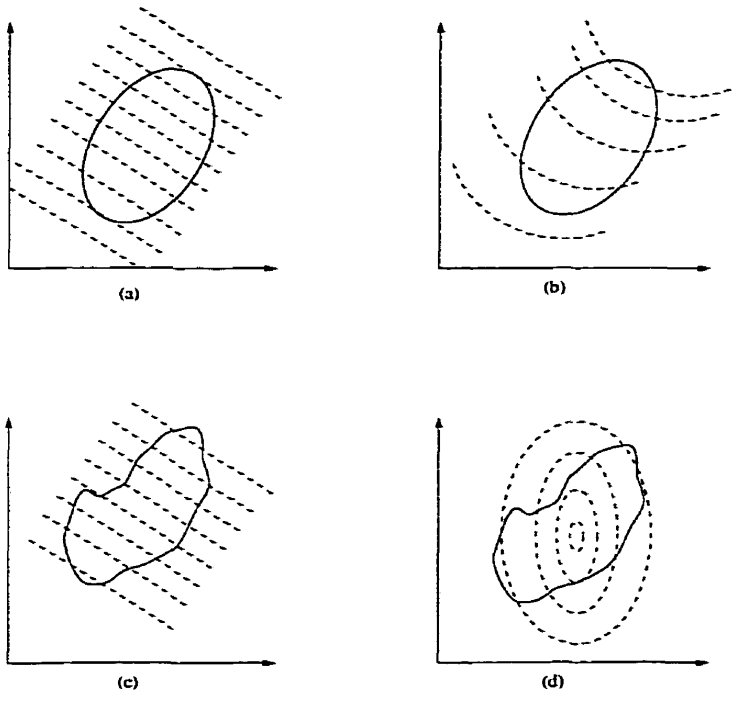

Figure 2.5: Graphical representations of inferences for various functions $g(\boldsymbol{\theta})$. Key: – likelihood region; - - contours of $g(\boldsymbol{\theta})$.

likelihood intervals for model predictions are always well approximated by linearization results even when the parameters of the model behave nonlinearly. It is easy to demonstrate that this is not true in general (see Chapter 3). There is no known reliable way to predict *a priori* whether or not a function $g(\boldsymbol{\theta})$ will behave nonlinearly. We postulate that nonlinearity in the behavior of $g(\boldsymbol{\theta})$ depends on the curvature of the contours of $g(\boldsymbol{\theta})$ *relative* to the curvature of the contours of the solution surface. For example, although the problem illustrated in Figure 2.6 is clearly nonlinear, it is likely that the linear approximation results would provide a good approximation to the true likelihood interval for $g(\boldsymbol{\theta})$ because the curvature of $g(\boldsymbol{\theta})$ is aligned with the curvature of $f(\boldsymbol{x}, \boldsymbol{\theta})$. There is still work to be done to explore this idea further. It may be possible to predict nonlinearity in $g(\boldsymbol{\theta})$ based on an *a priori* geometrical analysis of both $g(\boldsymbol{\theta})$ and $f(\boldsymbol{x}, \boldsymbol{\theta})$.

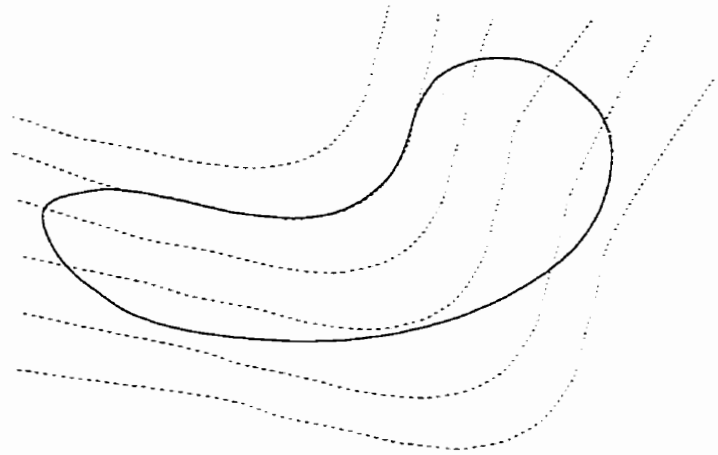

Figure 2.6: A Case of Nonlinear $g(\boldsymbol{\theta})$ and Nonlinear $f(\boldsymbol{x}, \boldsymbol{\theta})$.

2.4 Alternatives to Profiling

The most common approach to determining likelihood intervals for functions of parameters of nonlinear models is the linearization approach. This method is based on approximating the nonlinear model and the function of parameters $g(\boldsymbol{\theta})$ using a first order Taylor Series approximation centered at $\hat{\boldsymbol{\theta}}$. Then, the well-known linear inference results are applied. Using his approach, the likelihood interval for any function $g(\boldsymbol{\theta})$ is

$$g(\hat{\boldsymbol{\theta}}) \pm s\, t(n-p;\alpha/2)\sqrt{\frac{\partial g}{\partial \boldsymbol{\theta}}(\boldsymbol{V}^T\boldsymbol{V})^{-1}\frac{\partial g}{\partial \boldsymbol{\theta}}} \tag{2.29}$$

where $\boldsymbol{V}$ is an $n \times p$ matrix with elements defined by $\boldsymbol{V}_{ij} = \left.\frac{\partial f(x_i,\theta)}{\partial \theta_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$, and s is an estimate of the variance of the random error.

Halperin (1963), Rao (1973) and Khorasani and Milliken (1982) proposed that if a $100(1-\alpha)\%$ confidence region for the parameters is available, then one way to obtain confidence limits for a function of these parameters is to find the maximum and minimum of the function with respect to the parameters over the confidence region of the parameters. The geometrical representation of this is the same as that for profiling (see Figure 2.4). This results in confidence limits for $g(\boldsymbol{\theta})$ which are conservative

because the confidence regions are typically based on the critical value $F(p, n - p; \alpha)$, whereas confidence intervals are more appropriately based on $\sqrt{F(1, n - p; \alpha)}$ (Donaldson and Schnabel, 1987).

Model predictions are the most common examples of functions of parameters, and have received the most attention. Several papers have focused on assessing uncertainties in model predictions, and many algorithms for computing confidence bands have been proposed. Khorasani and Milliken (1979) derived conservative confidence bands for linear models based on a $1 - \alpha$ confidence region for the parameters. Later, Khorasani (1982) considered the problem of computing simultaneous confidence bands for nonlinear regression models. A confidence band is the band produced around the response function $f(\boldsymbol{x}, \boldsymbol{\theta})$ by considering simultaneously the uncertainty in the model predictions at all values of $\boldsymbol{x}$. Clarke (1987) used many of the ideas underlying profiling to construct approximate confidence limits for a function of the parameters of a nonlinear model. The method involves reparameterizing the model so that the function of interest becomes a parameter of the model. This is what is done in the reparameterization approach to generalized profiling (see Figure A.2). However, Clarke then advocated the use of a series of approximations which simplify the calculations. Given the speed of modern computers, such simplification is no longer necessary and we advocate a full likelihood approach to estimating uncertainty. More recent work on confidence bands has focused on asymptotic and large sample properties (Cox and Ma, 1995; Young et al., 1997). The reliability of these results for small and moderately-sized samples is unknown.

For some of the functions of parameters considered in this work, profiling is not an appropriate algorithm by which to compute the likelihood intervals (see Chapters 3, 4 and 7). For these cases, we have used a modified version of the minimization/maximization approach. We choose to maximize and minimize the function $g(\boldsymbol{\theta})$ over a likelihood region for $\boldsymbol{\theta}$ defined based on quantiles of the student t distribution;

that is, the upper limit of the likelihood interval for $g(\boldsymbol{\theta})$ is the value of $g(\boldsymbol{\theta})$ at the solution of the following optimization problem:

Maximize

$$g(\boldsymbol{\theta}) \tag{2.30}$$

subject to

$$|\tau(\boldsymbol{\theta})| \leq t(n-p;\alpha/2) \tag{2.31}$$

The lower limit of the likelihood interval for $g(\boldsymbol{\theta})$ is obtained by minimizing $g(\boldsymbol{\theta})$ subject to the constraint in (2.31). This approach to finding likelihood intervals for $g(\boldsymbol{\theta})$ produces results equivalent to those obtained using profiling. However, we favor the profiling algorithm because it provides information which, when displayed graphically, provides evidence of the nonlinearity of the problem. Also, for highly nonlinear $L(\boldsymbol{\theta})$ or $g(\boldsymbol{\theta})$, convergence problems may be encountered when using the minimization/maximization approach. These convergence problems can be avoided in profiling by choosing small step sizes, and using the location of the solution to the optimization problem in the current iteration as the starting guess for the optimization problem in the next iteration.

Chen (1991) also considered the minimization/maximization approach. However, he chose to assume that the maximum and minimum of $g(\boldsymbol{\theta})$ usually lie on the boundary of the likelihood region for $\boldsymbol{\theta}$, and not within it. Therefore, he considered the constraint in (2.31) to be an equality constraint. In our work, we have carried out the search over the entire likelihood region for $\boldsymbol{\theta}$.

The linear approximation approach to inference fails to account for the intrinsic and parameter effects curvature. Hamilton et al. (1982) developed approximate inference regions based on a quadratic approximation to the solution surface. By the

use of a quadratic approximation, some account is taken of the intrinsic curvature; however, all parameter-effects nonlinearities are still neglected. Likelihood intervals account for both of these types of nonlinearity.

Resampling methods have been used to obtain likelihood intervals for parameters in nonlinear models (Alpen and Gelb 1990; Bolviken and Skovlund, 1996). These methods are based on finding empirical approximations to the distribution of the statistic of interest by repeatedly simulating the system and estimating the statistic. Such methods include: Monte Carlo simulation, and the Jackknife and Bootstrap procedures (Davison and Hinkley, 1997; Jun and Tu, 1995). These methods are attractive because they require no *a priori* knowledge or assumptions about the distribution of the statistic. Only knowledge of the distribution of the random error entering the system is required. However, these methods are not appropriate for everyday use since they are computationally intensive and the results are specific to the given model (and to the assumptions made about the disturbance), and therefore cannot be generalized to other systems. These resampling methods often find use in the validation of other statistical estimation procedures.

Some of the same critisisms can be made of profiling. Profiling requires knowledge of the distribution of the random error entering the system, and the results of profiling are specific to the given model. However, Profiling is typically not so computationally intensive as resampling methods because very good starting values for the parameters are available for each optimization problem that is solved as part of the profiling algorithm.

An alternative approach, closely related to the likelihood ratio approach, is that of Hartley (1964) (also known as the lack-of-fit method). This method provides exact confidence regions for the complete complement of parameters in a nonlinear model. For subsets of the parameters and functions thereof, the method is only approximate (Donaldson and Schabel, 1987). It is based on a ratio of two independent quadratic

forms which follows an F distribution. Let

$$\boldsymbol{P}(\boldsymbol{\theta}) = \boldsymbol{V}(\boldsymbol{\theta})[\boldsymbol{V}^T(\boldsymbol{\theta})\boldsymbol{V}(\boldsymbol{\theta})]^{-1}\boldsymbol{V}^T(\boldsymbol{\theta}) \tag{2.32}$$

where $\boldsymbol{V}(\boldsymbol{\theta})$ is the $n \times p$ matrix of derivatives of the model with respect to the parameter. This notation is used to emphasize that in the method $\boldsymbol{V}$ is evaluated at any set of parameter values $\boldsymbol{\theta}$, and is not necessarily evaluated only at $\hat{\boldsymbol{\theta}}$ as in the linearization approach. Now consider the quadratic forms $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$

$$\boldsymbol{Q}_1(\boldsymbol{\theta}) = \frac{\boldsymbol{e}^T(\boldsymbol{\theta})\boldsymbol{P}(\boldsymbol{\theta})\boldsymbol{e}(\boldsymbol{\theta})}{\sigma^2} \tag{2.33}$$

$$\boldsymbol{Q}_2(\boldsymbol{\theta}) = \frac{\boldsymbol{e}^T(\boldsymbol{\theta})(\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{\theta}))\boldsymbol{e}(\boldsymbol{\theta})}{\sigma^2} \tag{2.34}$$

where $\boldsymbol{e}(\boldsymbol{\theta}) = \boldsymbol{y} - f(\boldsymbol{x}, \boldsymbol{\theta})$ is the residual vector. When $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, where $\boldsymbol{\theta}^*$ is the vector of true values of the parameters, $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ are independent χ^2 random variables with p and $n - p$ degrees of freedom, respectively. Therefore

$$\frac{\boldsymbol{Q}_1/p}{\boldsymbol{Q}_2/(n-p)} \sim F_{p,n-p} \tag{2.35}$$

and an exact $100(1-\alpha)\%$ confidence region for $\boldsymbol{\theta}$ is:

$$\frac{\boldsymbol{e}^T(\boldsymbol{\theta})\boldsymbol{P}(\boldsymbol{\theta})\boldsymbol{e}(\boldsymbol{\theta})}{\boldsymbol{e}^T(\boldsymbol{\theta})(\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{\theta}))\boldsymbol{e}(\boldsymbol{\theta})} \leq \frac{p}{n-p} F_{p,n-p,1-\alpha} \tag{2.36}$$

(Hamilton et al., 1982). The lack-of-fit method requires that $\boldsymbol{V}(\boldsymbol{\theta})$ be evaluated at a sufficient number of points $\boldsymbol{\theta}$ to produce a contour. Since, in the case of nonlinear regression models, likelihood ratio methods such as profiling require only that $f(\boldsymbol{x}, \boldsymbol{\theta})$ be evaluated at each point, the lack-of-fit method is more computationally intensive than profiling (Donaldson and Schnabel, 1987). This method does have an instructive geometrical interpretation. $\boldsymbol{P}(\boldsymbol{\theta})$ is the matrix of projection of $\boldsymbol{e}(\boldsymbol{\theta})$ onto the plane

tangent to $f(\boldsymbol{x}, \boldsymbol{\theta})$ at $\boldsymbol{\theta}$. Therefore, $\boldsymbol{e}^T(\boldsymbol{\theta})\boldsymbol{P}(\boldsymbol{\theta})\boldsymbol{e}(\boldsymbol{\theta})$ is the projection of the residual vector onto the tangent plane at $\boldsymbol{\theta}$ and $\boldsymbol{e}^T(\boldsymbol{\theta})(\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{\theta}))\boldsymbol{e}(\boldsymbol{\theta})$ is the projection of the residual vector onto a subspace orthogonal to the tangent plane at $\boldsymbol{\theta}$. At the true value of $\boldsymbol{\theta}$, these two projections are independent, and it is on this basis that their ratio follows an F distribution (Hamilton et al., 1982).

Prior to the work of Chen (1991) and Chen and Jennrich (1996), others had considered the problem of inference about functions of parameters in nonlinear regression. Some of the earliest work was done by Halperin and Mantel (1963) and Halperin (1964, 1965). In this series of papers, the problem of likelihood intervals was considered for a function $g(\mu_1, \mu_2)$, where μ_1 and μ_2 are unknown means of a bivariate normal distribution estimated from observations x_1 and x_2. The interval was based on maximizing and minimizing the function over a likelihood region for (μ_1, μ_2). Different approaches to determining the likelihood region for (μ_1, μ_2) were investigated, but the focus was on calibrating the value k so that the intervals based on $-2\ln(LR) \geq k$ would have coverage probabilities close to the nominal values.

Cook and Weisberg (1990) proposed a graphical approach to determining likelihood intervals for individual parameters in nonlinear regression. The proposed representation was simply the profile t plot with the τ statistic plotted along the x-axis and the parameter values plotted along the y-axis. The original work of Bates and Watts (1988) had the τ statistic plotted along the y-axis and the δ statistic (see (2.20) plotted along the x-axis. Of course the δ statistic is a linear function of the parameter values and so the profile t plot and the graph of Cook and Weisberg are equivalent. In this thesis, the profile t plots are shown as functions of the parameter values since this is deemed desirable when the intent is to obtain likelihood intervals for the parameters or functions of the parameters. Cook and Weisberg noted that such plots may be used to obtain likelihood intervals for model predictions if the model is reparameterized such that one of the parameters in the reparameterized model is defined

to be the prediction of interest; however they did not demonstrate this application.

Some nonlinear models are classified as partially linear (Knowels et al., 1991). These models have only one parameter that enters the model in a nonlinear way. Several inference results for such models have been reported in the literature. Williams (1962) developed an expression for an exact likelihood interval for the nonlinear parameter of a partially linear model. Later, Halperin (1963) extended this result to obtain an exact joint likelihood region for all of the parameters. Others, including: El-Saarrawi and Shah (1980) and Knowles et al. (1991) have further extended these early results. However, all of the inference results quoted in these works are relevant only to a very special class of models and are not generally relevant to inference problems in process control.

2.4.1 Robust Control Approaches to Uncertainty

There is a vast literature on robust control theory. The methods and theory in this area assume knowledge of the system model, complete with error bounds. In practice, the form of the model, the parameter values and the uncertainty bounds are all unknown and only estimates of these, based on available process data, can be obtained. Although the field of system identification has provided reliable tools and procedures for fitting and validating appropriate system models, there remains significant work to be done in the area of uncertainty estimation.

In system identification for robust control there are two competing philosophies. The first advocates a worst case approach to uncertainty. These so called hard bounding methods reject the traditional probabilistic approach to uncertainty in favor of bounds which are sure to be satisfied (i.e., which are hard). Common hard bounding algorithms include: the "unknown-but-bounded noise" algorithm, the "ellipsoidal" algorithm and "set membership" algorithms (Goodwin et al., 1992). Wahlberg and Ljung (1992) used set membership theory with a geometrical justification to develop

hard error bounds for linear transfer function models with bounded noise. The contributions to the error bounds by: noise, transient effects due to unmodeled disturbances and modeling error due to model/system mismatch, were all explicitly identified. Other work on developing algorithms for hard error bounds include contributions by Canale et al. (1998), Giarre et al. (1998), Gunnarson (1993) and Zhu (1989), among others. Regardless of the method, the hard bounding approaches result in error bounds which are highly conservative. They are necessarily so to ensure that even the worst case is enclosed by the bounds. Often, the bounds are unnecessarily conservative because conservative parameter space bounds can become even more conservative when transformed to transfer function space if the transformations are based on an approximation to the true expectation surface defined by $f(\boldsymbol{x}, \boldsymbol{\theta})$.

The hard bounding approaches are often advocated on the basis that robust control theory requires strictly true and known error bounds. However, for real systems, there is always uncertainty about the model and the disturbances entering the system. Therefore, it is unrealistic to propose methods to determine certain error bounds. To achieve near certainty, hard bounding methods overestimate the uncertainty and result in error bounds which are inappropriately wide. Goodwin et al. (1992) argued that hard error bounds are inappropriate because prior assumptions about noise, disturbances and control actions can never be known absolutely, and so the idea of certain limits is misguided. By this argument, a probabilistic approach is more consistent with the realities of system identification.

The second philosophy for estimating uncertainty in transfer function models is that of soft error bounds. It is based on a classical probabilistic approach. The soft bounds are not guaranteed to contain the system performance, but rather, are said to have a specified probability of containing the system performance. A soft approach to estimating error bounds for transfer function models was developed by Goodwin et al. (1992). They considered error due to bias from model/system mismatch, and

error due to noise in the measured data. The bias error was estimated by assuming a stochastic prior model for the distribution of the unmodeled dynamics. This model was embedded into the system estimation problem so that its parameters were estimated along with the parameters of the system model. These ideas were extended by Schoukens and Pintelon (1994) to allow for the case of colored noise in continuous-time systems explored in the frequency domain. Like most work to date in this area, the method is restricted to models which are linear in the parameters. Goodwin et al. (1992) did apply their method to an autoregressive moving average model with exogenous input variables (ARMAX) model by first linearizing the model. The authors claimed that their method performed well even for this case; however, there is obvious room for improvement in the case of nonlinear models.

DeVries and Van den Hof (1995) adopted a mixed deterministic/probabilistic approach to determining error bounds accounting for three sources of uncertainty: undermodeling, noise disturbance and unknown initial conditions. The model error due to unmodeled dynamics and unknown past input signals were considered deterministic worst-case quantities, and the noise disturbance was considered to be stochastic. Their algorithm was a two-step approach whereby the bias was estimated first, and the noise uncertainty was estimated second. Again, only linear finite impulse response (FIR) models were considered.

Ninness and Goodwin (1995a) examined the relationship between bounded-error and stochastic estimation theory, and showed that for many problems, the two approaches are equivalent when a Bayesian framework is used. A good overview of estimating uncertainty in models used for control is given by Ninness and Goodwin (1995).

2.5 Measures of Nonlinearity

A topic intimately related to reliable inference about $g(\boldsymbol{\theta})$ is the idea of measures of nonlinearity. A measure of nonlinearity is intended to be an indicator of the degree of nonlinearity of the problem, and is used to judge whether or not linearization inference results are likely to be reliable.

Most of the work in this area has focused on measuring the "average" or maximum curvature of the expectation surface (Beale, 1960; Bates and Watts, 1980). However, some work has been done on quantifying the nonlinearity of subsets of parameters (Cook and Goldberg, 1986; Clarke, 1987b), and recently, of functions of parameters (Kang and Rawlings, 1998). The measures of nonlinearity of Bates and Watts were developed for use with nonlinear regression models. Since then, Ravishanker (1994) has extended these ideas to the domain of time series analysis.

The measures of nonlinearity previously proposed in the literature have complicated expressions, and require that second derivatives of $f(\boldsymbol{x}, \boldsymbol{\theta})$ and $g(\boldsymbol{\theta})$ with respect to the parameters be found. Furthermore, the measures of Bates and Watts (1980) have been found to be unreliable in some cases (Cook and Witmer, 1985; van Ewijk and Hoekstra, 1994). There is a need for a simple yet reliable indicator of nonlinearity for time series models. Detailed background on the measures of nonlinearity proposed by Bates and Watts (1980) and Kang and Rawlings (1998) is given in Chapter 6. Here we simply note that these measures use a quadratic approximation of the model to compute an average nonlinearity and a maximum nonlinearity. The methods separate intrinsic nonlinearity from parameter-effects nonlinearity on the basis of the quadratic approximation to the expectation surface.

A new measure of nonlinearity for ARMA time series models is defined in Chapter 6. It is based on the fact that the nonlinearity of the parameters of these models is a function of the proximity of the parameter vector to the nearest stability/invertibility boundary.

2.6 Expected Profiling

Chapter 5 outlines a new methodology called expected profiling. Whereas generalized profiling and the measures of nonlinearity discussed so far are data driven (i.e., their values depend on a set of measured data), expected profiling is a tool for use in the absence of data (i.e., it does not depend on a specific set of data). Expected profiling is particularly appropriate for ARMA models, and has applications in predicting nonlinearity and designing experiments.

In the case of ARMA models, the model itself defines how the observations are likely to evolve over time. This feature allows us to derive an expected value for τ^2, which depends on the form of the model, the values of the parameters, and n, the length of the time series. We exploit the dependence on n to develop a methodology for use in determining how many observations will be required to achieve acceptable precision in the parameters in a proposed model.

Prior work on designing dynamic experiments has focused on the issue of choosing an input signal having optimal frequency characteristics for the estimation of transfer function models (Zarrop, 1979; Ljung, 1987; Shirt et al., 1994). The length of the input sequence must also be chosen, but little guidance has been provided concerning how much data must be collected in order to estimate a dynamic model reliably.

These ideas also raise the issue of how much data is required for asymptotic results to apply. There is an abundance of literature on the asymptotic properties of ARMA models and their parameters (Taniguchi, 1986; Frydman, 1980; Ljung, 1985; Zhu, 1989). These results apply when n is "large", but few authors state how large n should be. In Chapter 5, expected profiling is used to generate an n-plot, which shows how the likelihood limits for $g(\boldsymbol{\theta})$ are expected to change as the value of n changes. For comparison, the expected confidence intervals based on the Cramer Rao lower bounds are also shown on the same graph. The degree of agreement between the limits of the intervals based on expected profiling and those based on the Cramer Rao

lower bounds provides a qualitative measure of how reliable the asymptotic results are likely to be for a given value of n.

Expected profiling may also be used to plot expected profile plots, which are analogous to profile t plots but not based on a specific data set. When a realization of a process becomes available, the expected profile can be compared to the profile t plot, and the difference between the two may be used as a measure the extent to which peculiarities of that particular data set are contributing to the uncertainty and nonlinearity of the estimation problem. Bates and Watts (1980), along with others, developed measures of nonlinearity to separate nonlinearity due to parameterization from nonlinearity due to the form of the model. Expected profiling may be used as a tool to separate nonlinearity due to the data itself from nonlinearity due to a proposed model and its parameterization.

Although the manuscript in each of the succeeding chapters is a complete paper, and includes discussion of relevant computational analyses, Appendix A provides explicit details about the algorithms that are used.

The manuscripts which make up this thesis all have at their core the idea of profiling. Nonetheless, their subject areas are quite diverse. The papers touch on issues of inference, measuring nonlinearity and design of experiments. The models considered range from nonlinear regression models to discrete dynamic transfer function models. Contributions are made to the disciplines of applied statistics, chemical engineering and control theory.

2.7 Nomenclature

a_t	= white noise sequence
$a_1, a_2, \ldots, a_{na}$	= coefficients of the $A(q^{-1})$ polynomial
$\boldsymbol{a}$	= $p \times 1$ vector of constants
$A(q^{-1})$	= polynomial in the backshift operator q^{-1}
$b_1, b_2, \ldots, b_{nab}$	= coefficients of the $B(q^{-1})$ polynomial
$B(q^{-1})$	= polynomial in the backshift operator q^{-1}
c	= a constant
$c_1, c_2, \ldots, c_{nc}$	= coefficients of the $C(q^{-1})$ polynomial
$C(q^{-1})$	= polynomial in the backshift operator q^{-1}
d	= delay between a change in u_t and its effect on y_t
$\boldsymbol{e}$	= $n \times 1$ column vector of estimated random errors
$F(p, n-p; \alpha)$	= upper α quantile for the F distribution with p and $n-p$ degrees of freedom
$g(\boldsymbol{\theta})$	a function of parameters
$L(\boldsymbol{\theta})$	= likelihood function evaluated at $\boldsymbol{\theta}$
$LI(g(\boldsymbol{\theta}))$	= likelihood interval for $g(\boldsymbol{\theta})$
LR	= likelihood region
n	= number of observations
na, nb, nc	= orders of the polynomials $A(q^{-1})$, $B(q^{-1})$, $C(q^{-1})$, respectively
p	= number of estimated parameters
s	= estimated standard deviation of the random errors
$S(\boldsymbol{\theta})$	= sum of squared errors
se	= standard error

$t(n-p;\alpha/2)$ = upper $\alpha/2$ quantile of the t distribution with $n-p$ degrees of freedom

u_t = input to the process at time t

$\boldsymbol{V}$ = $n \times p$ matrix of elements v_{ij} representing the first derivative of $f(\boldsymbol{x}_i, \boldsymbol{\theta})$ with respect to the j^{th} parameter

$\boldsymbol{x}$ = $1 \times m$ row vector of m independent variables

$\boldsymbol{X}$ = $n \times p$ matrix of elements x_{ij} representing the level of the j^{th} independent variable for observation i

y = response variable

$\boldsymbol{y}$ = $n \times 1$ column vector of values of the response variable

Greek letters

α = significance level

$\delta(g(\boldsymbol{\theta}))$ = studentized value of $g(\boldsymbol{\theta})$

Δ = a scalar constant

ϵ = additive random error

$\boldsymbol{\epsilon}$ = $n \times 1$ column vector of random errors

θ_i = i^{th} parameter of a model

θ_l = the l^{th} parameter of the polynomial $\theta(q^{-1})$

$\boldsymbol{\theta}$ = $p \times 1$ vector of parameters

$\hat{\boldsymbol{\theta}}$ = $p \times 1$ vector of maximum likelihood estimates of the parameters

$\tilde{\boldsymbol{\theta}}$ = location of a constrained maximum of $L(\boldsymbol{\theta})$

$\theta(q^{-1})$ = moving average polynomial of a time series model

λ = a constant which defines the confidence level

σ	= standard deviation
$\tau(g(\boldsymbol{\theta}))$	= profile t statistic for $g(\boldsymbol{\theta})$
ϕ_k	= the k^{th} parameter of the polynomial $\phi(q^{-1})$
$\phi(q^{-1})$	= autoregressive polynomial of a time series model
$\chi^2(1)$	= the chi-squared distribution with 1 degree of freedom

Superscripts

$*$	= a true value
$\hat{}$	= a maximum likelihood estimate
$\tilde{}$	= a constrained estimate

Abbreviations

ARMA	autoregressive moving average
ARMAX	autoregressive moving average with exogenous inputs
ARX	autoregressive with exogenous inputs
FIR	finite impulse response
iid	independently and identically distributed
ML	maximum likelihood
MLE	maximum likelihood estimate

Chapter 3

Assessing the Precision of Model Predictions and Other Functions of Model Parameters

3.1 Abstract

Models fitted to data are used extensively in chemical engineering for a variety of purposes, including simulation, design and control. In any of these contexts it is important to assess the uncertainties in the estimated parameters and in any functions of these parameters, including predictions from the fitted model. Profiling is a likelihood ratio approach to estimating uncertainties in parameters and functions of parameters. A comparison is made between the optimization and reparameterization approaches to determining likelihood intervals for functions of parameters. The merits and limitations of generalized profiling are discussed in relation to the linearization approach commonly used in engineering. The benefits of generalized profiling are illustrated with two examples. A geometric interpretation of profiling is used to elucidate its value, and cases are identified for which the numerical algorithm fails.

3.2 Introduction

Models are used extensively in chemical engineering for simulation, design, control and testing of hypotheses of underlying phenomena. A model is usually "calibrated" by fitting it to experimental or process data. Before using a fitted model it is important to evaluate the uncertainty in any statistics of interest derived from the model. Such statistics may include the parameters themselves or functions of parameters, such as model predictions of yields and compositions. Occasionally, a model may be used to estimate some quantity or process characteristic which cannot be directly measured. For example, in process control the cross-over frequency may be of interest.

For the purpose of illustration, consider the series reaction

$$A \xrightarrow{\theta_1} B \xrightarrow{\theta_2} C$$

where chemical species A is consumed to produce species B which is subsequently consumed in a reaction producing species C. Let y_1, y_2 and y_3 be the molar concentrations of chemical species A, B and C, respectively, present at time t. Given the initial conditions: $y_1 = 1$ and $y_2 = y_3 = 0$ at $t = 0$, and assuming first order kinetics, the model for the molar concentration of species B can be written as

$$y_2 = \frac{\theta_1}{\theta_1 - \theta_2}[exp(-\theta_2 t) - exp(-\theta_1 t)] \tag{3.1}$$

It may be important to know the time

$$t_{max} = \frac{1}{\theta_1 - \theta_2}[ln(\theta_1) - ln(\theta_2)] \tag{3.2}$$

at which the concentration of B reaches a maximum. Note that t_{max} is a function of the parameters of the model (3.1). But to make judicious use of the estimates of the parameters (θ_1, θ_2) or of this function of parameters (t_{max}), the uncertainty in each

of those estimates must also be quantified, for it is this uncertainty which prescribes how much trust should be placed in the estimate.

For models which are linear in the parameters, and for which the random errors in the observations are normally and independently distributed (the usual least squares assumptions), uncertainty in the parameters or in any linear combination of the parameters can be readily calculated. The results are analytic and exact. This uncertainty may be expressed via confidence intervals or confidence regions for the parameters and confidence intervals for linear functions of the parameters. (The term "inference result" will be used to refer to confidence intervals, confidence regions, likelihood intervals or likelihood regions). For models which are nonlinear in the parameters, hereafter called nonlinear models, inference results for the parameters, or for any function of the parameters, are complex functions of: the distribution of the random errors associated with the measured responses, the structure of the model, the parameterization of the model, and the design of the experiment (Donaldson and Schnabel, 1987; Bates and Watts, 1988). Exact inference results are generally unavailable for nonlinear models, except for special cases (see for example: Williams, 1962; Halperin, 1963; Roy, 1993). Typically, approximate inference results for nonlinear models have been based on linear approximations to the models. This linear approximation approach is attractive in that it provides computationally simple results. However, Bates and Watts (1980), Ratkowsky (1983), Donaldson and Schnabel (1987), and others have shown that linear approximations are often poor, rendering results that are unreliable and possibly misleading.

More reliable inference results for the parameters of a nonlinear model can be obtained using a technique called profiling (Bates and Watts, 1988; Lam and Watts, 1991; Chen, 1991; Severini and Staniswalis, 1994; Chen and Jennrich, 1996). This technique, pioneered by Bates and Watts (1988), is a graphical method for displaying inference results. Chemical engineering examples are discussed extensively in Bates

and Watts (1988); a tutorial on the method is given in Watts (1994).

The idea of extending profiling to compute inference results for functions of model parameters has been suggested by Bates and Watts (1988), Ratkowsky (1983), Clarke (1987), Ross (1990), and Chen (1991)). However, only recently have the appropriate theory and computational methods been developed (Clarke, 1987; Chen, 1991; Chen and Jennrich, 1996). Two apparently different approaches have been proposed: one is based on a reparameterization of the model and the other involves a constrained optimization. Both methods, though equivalent, are conceptually and computationally different.

The purpose of this paper is to provide a comprehensive analysis of these methods and the computational procedures for profiling functions of parameters. Our intent is to elucidate and consolidate the theory underlying both methods, to assess the relative advantages and limitations of each method, and to indicate the usefulness of these methods for two chemical engineering examples.

The paper proceeds as follows. We begin by briefly reviewing the use of likelihood methods to construct inference regions for parameters and functions of parameters. This is followed by a discussion of profiling approaches for inference. A comparison is then made between the optimization approach and the reparameterization approach for determining likelihood intervals for functions of parameters. Two chemical engineering examples are then used to illustrate the methods. Following this section, we discuss a number of specialized cases where profiling may fail. The paper concludes with a discussion of some outstanding issues.

3.3 Likelihood intervals and regions

In this paper we restrict our attention to observations which can be modeled as:

$$y = f(\boldsymbol{x}, \boldsymbol{\theta}) + \epsilon \tag{3.3}$$

where the function $f(\boldsymbol{x}, \boldsymbol{\theta})$ is the expected value of the response variable y at specified levels of m independent variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$ and specified values of p parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$. ϵ is an additive random error term associated with y. Many models can be manipulated so that they can be expressed in this form including time series models and models having multiplicative error. Equation (3.3) can be generalized for a vector of n observed values $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ of the response variable:

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}) + \boldsymbol{\epsilon} \tag{3.4}$$

where $\boldsymbol{X}$ is now an $n \times p$ matrix of elements $x_{i,j}$ representing the level of the j^{th} independent variable for observation i. Steady state, dynamic and time series models can be included in this framework. The profiling methods described in this paper can be applied to any situation where the joint distribution of the observations can be specified. In most cases, it is assumed that the random errors $\boldsymbol{\epsilon}$ of the measured response variable are independently and identically normally distributed with mean zero and variance σ^2.

Estimates of the parameters in model (3.3) are often chosen to minimize the sum of squares of residuals, which are the deviations of the observations from the predictions:

$$\begin{aligned} S(\boldsymbol{\theta}) &= \sum_{i=1}^{n} (y_i - \boldsymbol{f}(\boldsymbol{x}_i, \boldsymbol{\theta}))^2 \\ &= (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}))^T (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta})) \\ &= \boldsymbol{e}^T \boldsymbol{e} \end{aligned} \tag{3.5}$$

where $\boldsymbol{e}^T = (e_1, e_2, \ldots, e_n)$ is the vector of residuals. The values of the parameters which minimize (3.5) is denoted by $\hat{\boldsymbol{\theta}}$. When the model is nonlinear, a numerical optimization method must be used to find $\hat{\boldsymbol{\theta}}$. The choice of the sum of squares

objective function can be justified without regard to the distribution of the random errors. However, to make inferences about the parameters or any function of the parameters, such as a prediction, it is necessary to specify the distribution of the random errors. As shown later, for the error distribution assumed in this paper, minimizing the objective function (3.5) yields maximum likelihood estimates of the parameters.

For models which are linear in the parameters, i.e., $\boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}) = \boldsymbol{X}\boldsymbol{\theta}$, and in which the random errors $\boldsymbol{\epsilon}$ of the measured response are assumed to be independently and identically normally distributed with mean zero and variance σ^2, analytic expressions for confidence intervals and confidence regions for the parameters and any linear combination of the parameters are well known and readily available. To summarize (Bates and Watts, 1988):

1. The joint confidence region for $\boldsymbol{\theta}$ is defined by the ellipsoid:

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{X}^T \boldsymbol{X} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq p s^2 F(p, n-p; \alpha) \tag{3.6}$$

   where $s = \sqrt{\frac{S(\hat{\boldsymbol{\theta}})}{n-p}}$ is an estimate of the standard deviation of the random errors, $(1-\alpha)$ is the level of confidence, and $F(p, n-p, \alpha)$ is the upper α quantile of the F distribution with p and $n-p$ degrees of freedom.

2. The marginal or individual confidence interval for the parameter θ_q (the q^{th} parameter in the model) is:

$$\hat{\theta}_q \pm s t(n-p; \alpha/2) \sqrt{\{(\boldsymbol{X}^T \boldsymbol{X})^{-1}\}_{qq}} \tag{3.7}$$

   where $\hat{\theta}_q$ is the least squares estimate of θ_q, $t(n-p, \alpha/2)$ is the upper $\alpha/2$ quantile of the t distribution with $n-p$ degrees of freedom, and $[(\boldsymbol{X}^T \boldsymbol{X})^{-1}]_{qq}$ is the q^{th} diagonal entry in the inverse of $\boldsymbol{X}^T \boldsymbol{X}$.

3. The confidence interval for any linear function of the parameters $g(\boldsymbol{\theta}) = \boldsymbol{a}^T\boldsymbol{\theta}$ is:

$$g(\hat{\boldsymbol{\theta}}) \pm st(n-p;\alpha/2)\sqrt{\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{a}} \tag{3.8}$$

Note that the expected value of a prediction at $\boldsymbol{x}_k$ from a linear model is a linear function of the parameters $\boldsymbol{a}^T\boldsymbol{\theta} = \boldsymbol{x}_k\boldsymbol{\theta}$. Equation 3.8 provides a confidence interval for the expected (mean) value of y at $\boldsymbol{x}_k$. To obtain a confidence interval for a new observation of y at $\boldsymbol{x}_k$, the expression

$$y(\boldsymbol{x}_k, \hat{\boldsymbol{\theta}}) \pm st(n-p;\alpha/2)\sqrt{1+\boldsymbol{x}_k^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_k} \tag{3.9}$$

is used. This expression takes into account the variance of the random error that is expected to affect a new observation. These ideas are discussed in detail in Chapter 7. In this chapter, "a prediction" is always used to mean the expected value of y at a point.

When models are nonlinear in the parameters, approximate confidence intervals can be obtained by using a quadratic approximation to the sum of squares function at the least squares estimate. This is equivalent to using a linear approximation to the model; hence we refer to this approach to constructing inference results as the linear approximation approach. The corresponding inference approximations are:

1. The approximate joint confidence region for $\boldsymbol{\theta}$ is defined by the ellipsoid:

$$(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\boldsymbol{V}^T\boldsymbol{V}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) \leq ps^2F(p,n-p;\alpha) \tag{3.10}$$

where $\boldsymbol{V}$ is an $n \times p$ matrix with elements defined by $\boldsymbol{V}_{ij} = \left.\frac{\partial f(x_i,\theta)}{\partial \theta_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$.

2. The marginal or individual confidence interval for the parameter θ_q is:

$$\hat{\theta}_q \pm st(n-p;\alpha/2)\sqrt{\{(\boldsymbol{V}^T\boldsymbol{V})^{-1}\}_{qq}} \tag{3.11}$$

3. The confidence interval for any function of the parameters $g(\boldsymbol{\theta})$ is:

$$g(\hat{\boldsymbol{\theta}}) \pm st(n-p;\alpha/2)\sqrt{\dot{\boldsymbol{g}}^T(\boldsymbol{V}^T\boldsymbol{V})^{-1}\dot{\boldsymbol{g}}} \tag{3.12}$$

where $\dot{\boldsymbol{g}} = \frac{\partial g}{\partial \boldsymbol{\theta}}|_{\hat{\boldsymbol{\theta}}}$.

Although this approach provides a computationally simple means of finding inference intervals, it can be misleading for some nonlinear models because it does not account for the curvature of the expectation surface nor the nonlinearity of the mapping of $\boldsymbol{\theta}$ from the observation space to the parameter space (Ratkowsky, 1983; Donaldson and Schnabel, 1987; Bates and Watts, 1980).

To develop more accurate inference regions for nonlinear models, or for any model in which constraints on the parameters may apply, it is helpful to resort to a likelihood interpretation of the parameters (Box and Tiao, 1973). When the random errors of the measured response variable are assumed to be independently and identically normally distributed with mean zero and variance σ^2, the likelihood function of the parameters is (Eliason, 1993):

$$L(\boldsymbol{\theta},\sigma^2|\boldsymbol{y}) = (2\pi\sigma^2)^{-n/2}\exp\left(\frac{-\boldsymbol{e}^T\boldsymbol{e}}{2\sigma^2}\right) \tag{3.13}$$

$$= (2\pi\sigma^2)^{-n/2}\exp\left(\frac{-S(\boldsymbol{\theta})}{2\sigma^2}\right) \tag{3.14}$$

where $\boldsymbol{e}$ is the vector of residuals $(\boldsymbol{y} - \boldsymbol{y}_{measured})$ resulting from the values $\boldsymbol{\theta}$ for the parameters. Maximizing (3.14) is equivalent to minimizing $S(\boldsymbol{\theta})$ and therefore the maximum likelihood estimates of $\boldsymbol{\theta}$ are equal to the least squares estimates.

A $(1-\alpha)$ likelihood *interval* for θ_q is the set of all values of θ_q for which $f(\boldsymbol{x},\boldsymbol{\theta})$ lies within a fixed distance of $\boldsymbol{y}$, with the values of the remaining parameters in the fitted model fixed at those values which maximize the conditional likelihood function (Chen and Jennrich, 1996). This is equivalent to saying that the likelihood interval for θ_q includes all values of θ_q such that $L(\boldsymbol{\theta}) \leq k$, where $L(\boldsymbol{\theta})$ is the conditional likelihood function of the parameters using $s^2 = \frac{S(\hat{\boldsymbol{\theta}})}{n-p}$ as an estimate of σ^2, and k is a constant whose value depends on the distribution of $\boldsymbol{e}$, the confidence level $(1-\alpha)$, and the number of degrees of freedom, $n-p$ associated with s^2.

An exact likelihood *region* is defined by the boundary of a contour of $L(\boldsymbol{\theta})$; this region need not be closed. For the model form and the random error assumptions considered in this paper, this contour is defined by those combinations of parameter values which satisfy $S(\boldsymbol{\theta}) = constant$. Mathematically, a nominal $(1-\alpha)$ likelihood region for the parameters is defined by the region for which:

$$S(\boldsymbol{\theta}) = S(\hat{\boldsymbol{\theta}}) \left[1 + \frac{p}{n-p} F(p, n-p; \alpha)\right] \tag{3.15}$$

It is important to note that, for linear models, the likelihood region defined by this equation is exact both in terms of the shape of the region and the probability or confidence level. However, for nonlinear models, although the shape of the region will be exact, the confidence level will only be approximate (Draper and Smith, 1981). Statisticians speak of the coverage probability of the interval or region. This is a measure of the true confidence level of the inference and may be estimated by means of Monte Carlo simulations. This can require a very large computational effort before the estimated confidence level can be deemed acceptable.

Typically, the likelihood interval is based on a likelihood ratio approach to inference (Lehmann, 1959). Consider the null hypothesis $\theta_q = c$ versus the alternate

$\theta_q \neq c$. The likelihood ratio for testing this null hypothesis is:

$$LR(c) = \frac{L(\tilde{\boldsymbol{\theta}})}{L(\hat{\boldsymbol{\theta}})} \tag{3.16}$$

where $LR(c)$ is the likelihood ratio for the null hypothesis, $L(\hat{\boldsymbol{\theta}})$ is the likelihood function evaluated at $\hat{\boldsymbol{\theta}}$, the unconditional maximum likelihood estimates of all parameters in the model, $L(\tilde{\boldsymbol{\theta}})$ is the conditional maximum likelihood given that $\theta_q = c$, and $\tilde{\boldsymbol{\theta}}$ is the set of maximum likelihood estimates of the parameters conditional on $\theta_q = c$.

For the case of ordinary regression, where the errors are independently and identically normally distributed with mean zero and variance σ^2, the likelihood ratio is:

$$LR(c) = \exp\left(\frac{S(\hat{\boldsymbol{\theta}}) - S(\tilde{\boldsymbol{\theta}})}{2\sigma^2}\right) \tag{3.17}$$

where $S(\tilde{\boldsymbol{\theta}})$ is the minimum sum of squared residuals conditional on $\theta_q = c$. It can be shown that

$$-2 \ln LR(c) \sim \chi^2(1) \tag{3.18}$$

where $\chi^2(1)$ is the chi-squared distribution with one degree of freedom. For linear models, expression (3.18) can be manipulated to obtain the confidence interval given in (3.11).

Profiling was first developed as a numerical method to estimate reliable likelihood intervals for parameters of steady-state nonlinear models in which the random error entering the process is assumed to be normally distributed. The focus of this paper is inference about functions of parameters rather than inference about the parameters themselves. The following sections describe the development of the profiling algorithm in a very general framework so that it can be used to solve the inference problem of

finding likelihood intervals for a general function of parameters $g(\boldsymbol{\theta})$.

3.4 Profiling

The profiling algorithm was first proposed by Bates and Watts (1988) as a means of finding likelihood intervals for individual parameters of nonlinear models. Because the algorithm was derived for this specific purpose, it was developed in terms of sums of squares of residuals. The motivation for the original algorithm can be demonstrated by looking at the linear likelihood interval given in (3.7). This interval can be equivalently expressed as:

$$-t(n-p,\alpha/2) \leq \frac{\theta_q - \hat{\theta}_q}{se(\hat{\theta}_q)} \leq t(n-p,\alpha/2) \tag{3.19}$$

(Bates and Watts, 1988). That is, an exact $(1-\alpha)$ likelihood interval for the q^{th} parameter includes all values of θ_q that satisfy (3.19). Defining

$$\tau(\theta_q) = \frac{\theta_q - \hat{\theta}_q}{se(\hat{\theta}_q)} \tag{3.20}$$

(3.19) can be rewritten as:

$$-t(n-p,\alpha/2) \leq \tau(\theta_q) \leq t(n-p,\alpha/2) \tag{3.21}$$

$\tau(\theta_q)$ can also be defined as

$$\tau(\theta_q) = sign(\theta_q - \hat{\theta}_q)\sqrt{\frac{S(\tilde{\boldsymbol{\theta}}) - S(\hat{\boldsymbol{\theta}})}{s^2}} \tag{3.22}$$

(Bates and Watts, 1988).

Equations (3.20) and (3.22) are equivalent for linear models, but not for nonlinear models. Linearization inference results for nonlinear models are based on (3.20).

Profiling inference results are based on (3.22). The definition for $\tau(\theta_q)$ given in (3.22) does consider the curvature of the expectation surface.

In the profiling algorithm, as first proposed by Bates and Watts (1988), each parameter in turn is incremented in a series of small steps away from its least squares estimate. Conditional upon each new value of θ_q, the sum of squares of the residuals is minimized with respect to the remaining parameters. This determines $S(\bar{\boldsymbol{\theta}})$, and from the results of this optimization the $\tau(\theta_q)$ statistic is calculated. Typically, profiling results are presented in a profile t plot, which is a plot of τ versus θ_q. From Equation (3.20) it is obvious that the profile t plot for a linear model is always linear. For nonlinear models the profile t plot will be curved. The amount of curvature can be used as a qualitative means of assessing the degree of nonlinearity of the model with respect to the given parameter over the region of interest (Chen and Jennrich, 1996). The information accumulated through the profiling process can also be used to generate joint nominal likelihood regions for pairs of parameters in a way that is more computationally efficient than generating standard pairwise contour plots of the sum of squared residuals surface (Bates and Watts, 1988).

Equation (3.22) defines the τ statistic for any parameter of the model. However, the primary interest in this paper is in deriving profiling results for an arbitrary function g of the parameters, where

$$g = g(\boldsymbol{\theta}) \tag{3.23}$$

Since the profiling technique of Bates and Watts has its basis in likelihood ratio testing, the likelihood ratio formalism is an obvious starting point for extending the algorithm to a function of parameters $g(\boldsymbol{\theta})$. Chen and Jennrich (1996) defined the

statistics:

$$D^2(\boldsymbol{\theta}) = -2\ln\left(\frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})}\right) \tag{3.24}$$

and

$$z^*(c) = sign(c - \hat{g})D(\hat{\boldsymbol{\theta}}_c) \tag{3.25}$$

where $D^2(\boldsymbol{\theta})$ is called the deviance and $z^*(c)$ is called the signed root deviance (SRD), which is used as the basis for a profile plot called the signed root deviance profile (SRDP). Barndorff-Nielson (1986) defined a similar statistic. The SRDP has its basis in (3.17) and is therefore a generalization of the profile t plot of Bates and Watts. In this paper the likelihood ratio intervals and plots will be referred to as profile t plots to be consistent with that early work. The term *generalized profiling* will be used to denote the use of profile plots in a broader sense than originally described. Two approaches to generalized profiling have been proposed and they are developed in the following two sections.

3.5 Reparameterization

The first approach to generalizing the profiling algorithm is based on reparameterizing the model. This approach has been mentioned by several authors including Bates and Watts (1988), Ross (1990), and Watts (1994); however, no development of the theory was given. Clarke (1987) used reparameterization to develop an approximate algorithm for estimating confidence intervals. In this section we will present in detail the reparameterization approach to generalized profiling.

A new set of parameters $\boldsymbol{\phi} = (\phi_1, \phi_2, \cdots, \phi_p)^T$ is defined such that one of the new parameters is the function g of interest. In this way the standard profiling results can

be used directly. For example, the new parameters may be defined as:

$$\begin{aligned} \phi_1 &= g(\boldsymbol{\theta}) \\ \phi_2 &= \theta_2 \\ &\vdots \\ \phi_p &= \theta_p \end{aligned} \tag{3.26}$$

where $g(\boldsymbol{\theta})$ is the function of interest.

This reparameterization proceeds by solving the set of equations given in (3.26) for $\boldsymbol{\theta}$, such that

$$\begin{aligned} \theta_1 &= g^{-1}(\boldsymbol{\phi}) \\ \theta_2 &= \phi_2 \\ &\vdots \\ \theta_p &= \phi_p \end{aligned} \tag{3.27}$$

where g^{-1} is the inverse function of g. These results are then substituted into (3.3), and the model becomes

$$y = f(\boldsymbol{x}, \boldsymbol{\phi}) + \epsilon \tag{3.28}$$

The profile t plot for each of the new parameters can be found by computing the τ statistic, as defined in (3.22) for the case where $\epsilon \sim N(0, \sigma^2)$, or as defined in (3.25) for the general case, for each of the parameters in the reparameterized model. Because the parameter ϕ_1 was defined to be the function of interest, the profile t plot for ϕ_1 is the profile t plot for g, and the task at hand has been accomplished. The τ

statistic for the function g is given by:

$$\begin{aligned} \tau_g &= \tau_{\phi_1} \\ &= sign(\phi_1 - \hat{\phi}_1)\frac{\sqrt{S(\tilde{\boldsymbol{\phi}}) - S(\hat{\boldsymbol{\phi}})}}{s} \\ &= sign(g_c - \hat{g})\frac{\sqrt{S(\tilde{\boldsymbol{\phi}}) - S(\hat{\boldsymbol{\theta}})}}{s} \end{aligned} \tag{3.29}$$

for the case of independently normally distributed errors, where g_c is the specified value of ϕ_1 and $S(\tilde{\boldsymbol{\phi}})$ is the minimum sum of squared residuals, conditional on $g = g_c$. As in the original algorithm, the profile t plot for g is produced by incrementing g in a series of small steps above and below its least squares estimate, and calculating the conditional sum of squared residuals, $S(\tilde{\boldsymbol{\phi}})$, at each step. In developing Equation (3.29), use was made of the result that

$$\begin{aligned} \hat{\phi}_1 &= g(\hat{\boldsymbol{\theta}}) \\ \hat{\phi}_2 &= \hat{\theta}_2 \\ &\vdots \\ \hat{\phi}_p &= \hat{\theta}_p \end{aligned}$$

so that

$$S(\hat{\boldsymbol{\phi}}) = S(\hat{\boldsymbol{\theta}}) \tag{3.30}$$

and, therefore, s is invariant under reparameterization (Chen, 1991).

For the general case,

$$\tau_g = \tau_{\phi_1}$$

$$
\begin{aligned}
&= sign(\phi_1 - \hat{\phi}_1)\sqrt{-2\ln\frac{L(\tilde{\boldsymbol{\phi}})}{L(\hat{\boldsymbol{\theta}})}} \\
&= sign(c - \hat{g})\sqrt{-2\ln\frac{L(\tilde{\boldsymbol{\phi}})}{L(\hat{\boldsymbol{\theta}})}}
\end{aligned} \tag{3.31}
$$

In profiling, a reference line is constructed against which the nonlinearity of the model with respect to the given parameter can be evaluated. For the case where ϵ is normally distributed this reference line is based on (3.20), and it represents the profile t plot for the linear approximation to the model. When profiling a function g, the reference line is:

$$
\begin{aligned}
\tau_{lin}(g) &= \tau_{lin}(\phi_1) \\
&= \frac{\phi_1 - \hat{\phi}_1}{se(\hat{\phi}_1)} \\
&= \frac{c - \hat{g}}{se(\hat{g})}
\end{aligned} \tag{3.32}
$$

The standard error of $\hat{g}$, $se(\hat{g})$, is given by:

$$
\begin{aligned}
se(\hat{g}) &= se(\hat{\phi}_1) \\
&= s\sqrt{\left[(\boldsymbol{V}^T_{f(\hat{\boldsymbol{\phi}})}\boldsymbol{V}_{f(\hat{\boldsymbol{\phi}})})^{-1}\right]_{11}}
\end{aligned} \tag{3.33}
$$

where $\boldsymbol{V}_{f(\hat{\boldsymbol{\phi}})}$ is the $n \times p$ matrix with elements defined by $\boldsymbol{V}_{f(\hat{\boldsymbol{\phi}}),ij} = \left.\frac{\partial \boldsymbol{f}(x_i,\boldsymbol{\phi})}{\partial \phi_j}\right|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}}$. The subscript $f(\hat{\boldsymbol{\phi}})$ is used to emphasize that $\boldsymbol{V}$ represents the matrix of first derivatives of the *reparameterized* model.

A step-by-step algorithm for the reparameterization approach to generalized profiling is given in Figure (3.1).

1. Using a nonlinear optimization package, find the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$.

2. Compute the MLE of $g(\boldsymbol{\theta})$, and define $\hat{g} = g(\hat{\boldsymbol{\theta}})$.

3. Compute an estimate of the variance of the measurement error (i.e., compute s^2).

4. Compute $Cov(\hat{\boldsymbol{\theta}})$.

5. Compute $se(\hat{g}) = \sqrt{s^2 \left.\frac{dg}{d\boldsymbol{\theta}}^T\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{V}^T\boldsymbol{V})^{-1} \left.\frac{dg}{d\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}}$.

6. Reparameterize the model such that the first new parameter ϕ_1 is $\phi_1 = g(\boldsymbol{\theta})$.

7. Set the index i to 1, and let $g_{old} = g_{hat}$

8. Move the value of ϕ_1 away from $\hat{g}$ by a small amount Δ (i.e., $g_i = g_{old} + \Delta$) .

9. Create a new vector of $p-1$ parameters which does not include ϕ_1. Let the new vector be $\boldsymbol{\phi}_{reduced}$.

10. Use an unconstrained optimization package to solve the $(p-1)$-dimensional optimization problem. The location of the optimum is $\hat{\boldsymbol{\phi}}_{reduced}$.

11. In p-dimensional space, the location of the optimum is $\tilde{\boldsymbol{\phi}} = [g_i, \tilde{\boldsymbol{\phi}}^T_{reduced}]$.

12. Compute

$$\tau_i = sign(g_i - \hat{\phi}_1)\sqrt{-2ln\left(\frac{L(\tilde{\boldsymbol{\phi}})}{L(\hat{\boldsymbol{\phi}})}\right)}$$

$$\delta_i = \frac{g_i - \hat{\phi}_1}{se(\hat{\phi}_1)}$$

13. Is $|\tau_i| \geq t(n-p, \alpha/2)$? If yes, continue. If no, set $g_{old} = g_i$, set $i = i+1$, and return to Step 8.

14. Is Δ negative? If yes, continue. If no, set $g_{old} = \hat{g}$, set $i = i+1$, let $\Delta = -\Delta$ and return to Step 8.

15. Fit a smooth curve through g_i versus τ_i and use this to find the values of $g(\boldsymbol{\theta})$ at $\tau = \pm t(n-p, \alpha/2)$. These are the limits of the likelihood interval.

16. Compute the limits of the linearization confidence interval using

$$CI = \hat{g} \pm se(\hat{g})t(n-p, \alpha/2)$$

17. Construct the profile t plot by plotting, on one figure, τ_i versus g_i, and δ_i versus g_i.

Figure 3.1: A step-by-step algorithm for the reparameterization approach to generalized profiling.

3.6 The constrained optimization approach

Chen (1991) also developed a method for obtaining profiling results for an arbitrary function $g(\boldsymbol{\theta})$. He began by noting that when profiling a parameter θ_q, the following constrained optimization problem is being solved at each iteration of the algorithm:

Maximize

$$L(\boldsymbol{\theta}) \tag{3.34}$$

subject to the constraint

$$\theta_q = c$$

where c is a constant. The maximum is $L(\tilde{\boldsymbol{\theta}})$. Chen then asserted that the constraint in this maximization problem need not involve only one parameter. Rather, the constraint could be chosen to involve a function of the parameters. Profiling results for a function of the parameters, g, may then be obtained by solving the following constrained optimization problem:

Maximize

$$L(\boldsymbol{\theta}) \tag{3.35}$$

subject to the constraint

$$g(\boldsymbol{\theta}) = c$$

for a series of values of c above and below the least squares estimate of g.

Furthermore, Chen (1991) showed that the linearization likelihood interval for g

(assuming $\epsilon \sim N(0, \sigma^2)$) is:

$$g(\boldsymbol{\theta}) = g(\hat{\boldsymbol{\theta}}) \pm se(\hat{g})\, t(n-p; \alpha/2) \tag{3.36}$$

where

$$\{se(\hat{g})\}^2 = s^2 \left.\frac{dg}{d\boldsymbol{\theta}}^T\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{V}_{f(\hat{\boldsymbol{\theta}})}^T \boldsymbol{V}_{f(\hat{\boldsymbol{\theta}})})^{-1} \left.\frac{dg}{d\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{3.37}$$

and $\boldsymbol{V}_{f(\hat{\boldsymbol{\theta}})}$ is the $n \times p$ matrix with elements defined by $\boldsymbol{V}_{f(\hat{\boldsymbol{\theta}}),ij} = \left.\frac{\partial f(x_i,\boldsymbol{\theta})}{\partial \theta_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$.

A step-by-step algorithm for the optimization approach to generalized profiling is given in Figure (3.2).

Although the two approaches are equivalent (Quinn et al., 1999b; Chapter 4), the performance of the profiling algorithm, in terms of computation time and convergence, may depend on which approach is used in the implementation. Our experience with a number of examples indicates that neither approach is consistently better than the other in terms of performance. For example, the reparameterization method appears to perform better in cases where the parameter effects curvature (Bates and Watts, 1980) of the reparameterized model is lower than that of the original model. It should be noted that, for some applications, $g(\boldsymbol{\theta})$ may be a sufficiently complicated function of the parameters that it is difficult or impossible to find an analytic solution to the reparameterization of $f(\boldsymbol{x}, \boldsymbol{\theta})$. In such cases Chen's optimization approach is recommended. Although it is possible to use the reparameterization approach by performing the reparameterization numerically, our experience with the two algorithms has shown Chen's algorithm to converge faster and more consistently in such cases. These observations on the performances of the two profiling algorithms are strictly empirical. Further work to establish the topology of the solution surface for which one approach outperforms the other is required.

1. Using a nonlinear optimization package, find the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$.

2. Compute the MLE of $g(\boldsymbol{\theta})$, and define $\tilde{g} = g(\tilde{\boldsymbol{\theta}})$.

3. Compute an estimate of the variance of the measurement error (i.e., compute s^2).

4. Compute $Cov(\tilde{\boldsymbol{\theta}})$.

5. Compute $se(\hat{g}) = \sqrt{s^2 \left.\frac{dg}{d\boldsymbol{\theta}}^T\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{V}^T\boldsymbol{V})^{-1} \left.\frac{dg}{d\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}}$.

6. Set the index i to 1, and let $g_{old} = \tilde{g}$.

7. Move the value of $g(\boldsymbol{\theta})$ away from $\tilde{g}$ by a small amount Δ (i.e., $g_i = g_{old} + \Delta$).

8. Use a constrained nonlinear optimization package to solve the constrained optimization problem, maximize $L(\boldsymbol{\theta})$ subject to $g(\boldsymbol{\theta}) = g_i$. The location of the constrained optimum is $\tilde{\boldsymbol{\theta}}$.

9. Compute

$$\tau_i = sign(g_i - \tilde{g}) \sqrt{-2ln\left(\frac{L(\tilde{\boldsymbol{\theta}})}{L(\hat{\boldsymbol{\theta}})}\right)}$$

$$\delta_i = \frac{g_i - \tilde{g}}{se(\hat{g})}$$

10. Is $|\tau_i| \geq t(n-p, \alpha/2)$? If yes, continue. If no, set $g_{old} = g_i$, set $i = i + 1$, and return to Step 7.

11. Is Δ negative? If yes, continue. If no, set $g_{old} = \tilde{g}$, set $i = i + 1$, let $\Delta = -\Delta$ and return to Step 7.

12. Fit a smooth curve through g_i versus τ_i and use this to find the values of $g(\boldsymbol{\theta})$ at $\tau = \pm t(n-p, \alpha/2)$. These are the limits of the likelihood interval.

13. Compute the limits of the linearization confidence interval using

$$CI = \tilde{g} \pm se(\tilde{g}) t(n-p, \alpha/2)$$

14. Construct the profile t plot by plotting, on one figure, τ_i versus g_i, and δ_i versus g_i.

Figure 3.2: A step-by-step algorithm for the optimization approach to generalized profiling.

3.7 Geometric interpretation of profiling

A geometric interpretation of profiling is also possible. By definition, the nominal (1-α) likelihood interval for $g(\boldsymbol{\theta})$ contains all values of $g(\boldsymbol{\theta})$ such that:

$$\frac{S(\tilde{\boldsymbol{\theta}}) - S(\hat{\boldsymbol{\theta}})}{S(\hat{\boldsymbol{\theta}})} \leq \frac{1}{n-p} F(1, n-p; \alpha) \tag{3.38}$$

Geometrically, the limits of the interval correspond to the maximum and minimum values of $g(\boldsymbol{\theta})$ on or within the contour of $\boldsymbol{\theta}$ satisfying:

$$S(\boldsymbol{\theta}) \leq S(\hat{\boldsymbol{\theta}}) \left[1 + \frac{1}{n-p} F(1, n-p; \alpha)\right] \tag{3.39}$$

When both the model and $g(\boldsymbol{\theta})$ are linear, the solution to this constrained optimization problem is the analytic expression given in (3.8), and the confidence level is exact. We note that the contour satisfying (3.39) is always contained within the contour defined in (3.15) (i.e., the joint confidence region for the parameters) since it can be verified empirically that

$$pF(p, n-p; \alpha) \geq F(1, n-p; \alpha) \tag{3.40}$$

These ideas are illustrated in Figure 3.3 for the case of a two parameter linear model. In this case the maximum and minimum values of g occur at the points where the the contours of g are tangent to the ellipse defining the appropriate likelihood region for $\boldsymbol{\theta}$. When g and/or the model are nonlinear, there are several possible cases:

1. g_{max} and g_{min} occur on the boundary of the appropriate likelihood region for $\boldsymbol{\theta}$, as for the linear case.

2. g_{max} and/or g_{min} are/is located in the interior of the appropriate likelihood region for $\boldsymbol{\theta}$.

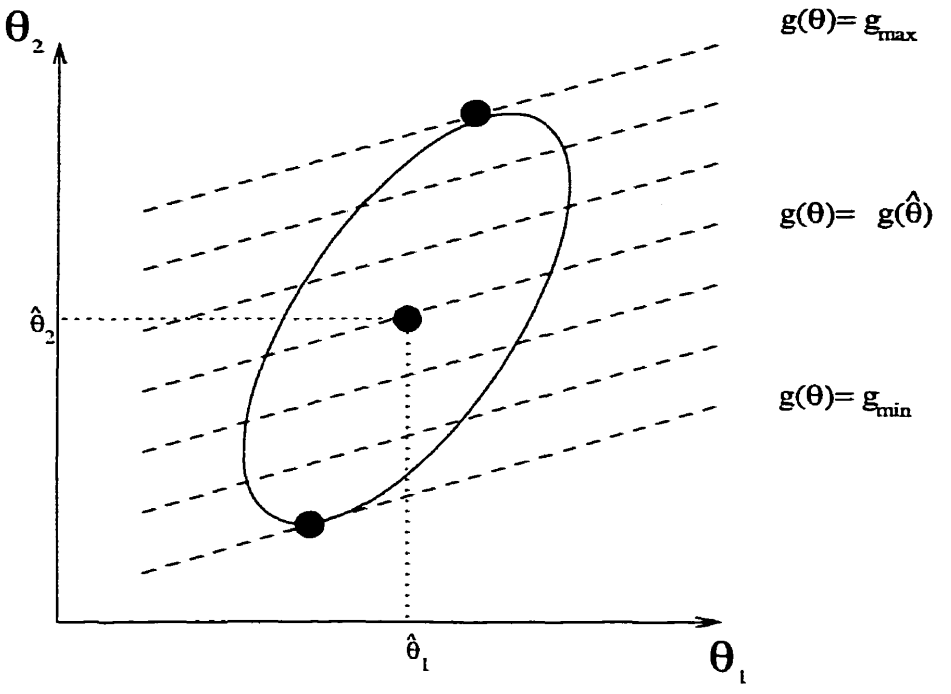

Figure 3.3: An illustration of the geometry of likelihood intervals for $g(\boldsymbol{\theta})$ when $g(\boldsymbol{\theta})$ and the model are linear functions of $\boldsymbol{\theta}$ (solid line: likelihood region; dashed lines: contours of $g(\boldsymbol{\theta})$).

3. g intersects the appropriate likelihood region for $\boldsymbol{\theta}$ at only a single point, and so the likelihood interval is a single point (i.e., there is only a single plausible value for g at the given level of confidence). Usually this implies that $g(\hat{\boldsymbol{\theta}})$ is undefined and that g is defined only at a single point which lies on the boundary of the likelihood region for $\boldsymbol{\theta}$.

4. g does not intersect with the appropriate likelihood region for $\boldsymbol{\theta}$. This case can occur if the parameters of the fitted model are not appropriately constrained during the fitting of the model. For example, a rate equation fit to noisy data from a poorly designed experiment may result in parameter estimates which are negative and likelihood regions which include only negative values. If, for example, the function $g(\boldsymbol{\theta}) = ln(\theta_1)$ is of interest, then g cannot be estimated. Of course, a negative rate constant is nonsensical, and the model should have been appropriately constrained before the fitting was performed. This constraining

can be accomplished by an appropriate reparameterization of the model.

5. g has multiple local optima which fall on or within the boundary of the likelihood region for $\boldsymbol{\theta}$. This may result in confidence intervals which are not continuous. (Quinn et al., 1999b; Chapter 4)

It is clear that nonlinearity significantly complicates the inference results. Often the likelihood intervals for a nonlinear $g(\boldsymbol{\theta})$ will be asymmetric about $g(\hat{\boldsymbol{\theta}})$, and for Case 5 the likelihood intervals may be disjoint (Donaldson and Schnabel, 1987).

3.8 Illustrative examples

3.8.1 Example 1

The isomerization example of Bates and Watts (1988) (based on the data of Carr, 1960) is used to illustrate the two approaches to generalized profiling. The data consist of 24 observations of the reaction rate of the catalytic isomerization of n-pentane to isopentane at known partial pressures of the reactants and products (i.e., of n-pentane, hydrogen and isopentane). The data are tabulated in Table 3.1. The model proposed by Carr for the reaction rate was a Hougen-Watson model of the form

$$y = \frac{\theta_1\theta_3(x_2 - x_3/1.632)}{1 + \theta_2 x_1 + \theta_1 x_2 + \theta_4 x_3} + \epsilon \tag{3.41}$$

where x_1 is the partial pressure of hydrogen, x_2 is the partial pressure of n-pentane, and x_3 is the partial pressure of isopentane. Note that the labeling of the parameters in this example differs slightly from that used by Bates and Watts (1988). The parameters θ_1 and θ_3 have been interchanged so that θ_1 in this illustration is equal to θ_3 in the Bates and Watts parameterization, and θ_3 here is equal to θ_1 in Bates and Watts. All subsequent references to the parameters in this example are consistent

with the parameterization used in (3.41). The parameters of this model have physical significance. Parameters θ_1, θ_2 and θ_4 are equilibrium adsorption constants, and θ_3 is a rate constant. It is important to keep in mind the physical significance of the parameters when interpreting the statistical results. Bates and Watts (1988) showed that the profile t plots for the parameters of this model (Figure 3.4) are drastically nonlinear and that there are nearly perfect correlations between the parameters θ_1, θ_2 and θ_4. However, Bates and Watts (1988) produced residual plots which do not indicate any significant lack of fit. The high degree of correlation among the parameters and much of the nonlinearity is likely due to the model being overparameterized relative to the information in the data. Because of the overparameterization not all of the parameters can be estimated; instead, only linear functions of θ_1, θ_2 and θ_4 may be estimated.

Now consider the case where the specific values of the parameters are not of interest, but the model is to be used for predictive purposes only. In this case the important profile t plot is that for the prediction of the expected value of the rate of reaction at specified partial pressures of n-pentane, isopentane and hydrogen. Suppose that the rate of reaction at $\boldsymbol{x} = (2069, 990.8, 621.9)$ is of special interest. Then:

$$g(\boldsymbol{\theta}) = \frac{\theta_1\theta_3(990.8 - 621.9/1.632)}{1 + \theta_2 2069 + \theta_1 990.8 + \theta_4 621.9} \tag{3.42}$$

Profile t plots for the parameters of the model and for $g(\theta)$ are shown in Figures 3.4 and 3.5, respectively. These plots were generated using the optimization approach to generalized profiling. A step-by-step algorithm for the optimization algorithm is given in Figure 3.2. All profiling results were generated using MATLAB 4.2c (The MathWorks, 1994). To profile the function of parameters given in (3.42) required only that the form of the model and of the function of parameters be specified along, with the maximum likelihood estimates of the parameters. The maximum likelihood

Table 3.1: Isomerization Data (Example 1).

x_1 Hydrogen (kPa) $\times 10^{-3}$	x_2 n-Pentane (kPa)$\times 10^{-3}$	x_3 Isopentane (kPa)$\times 10^{-3}$	y Reaction Rate (h^{-1})
1.42	0.63	0.26	3.541
2.79	0.64	0.25	2.397
1.45	1.21	0.34	6.694
2.77	1.29	0.31	4.722
1.55	0.64	0.80	0.593
2.78	0.70	0.89	0.268
1.47	1.39	0.93	2.797
2.80	1.33	0.93	2.451
0.92	0.97	0.60	3.196
3.25	0.99	0.60	2.021
2.07	0.47	0.56	0.896
2.08	1.48	0.70	5.084
2.05	0.98	0.07	5.686
2.16	1.01	1.83	1.193
2.11	0.98	0.59	2.648
2.07	0.99	0.62	3.303
2.11	0.97	0.60	3.054
2.10	0.98	0.60	3.302
2.07	0.57	0.46	1.271
0.74	1.45	0.23	11.648
2.88	0.58	0.23	2.002
1.73	2.03	0.29	9.604
1.73	1.02	0.10	7.754
1.00	2.01	0.35	11.590

estimates for the parameters and the predictions, and the corresponding inference results, are given in Table 3.2. The constrained optimization routine provided in the "Optimization Toolbox" for MATLAB was used to solve the constrained optimization problem which is the basis of each iteration of the profiling algorithm.

To profile $g(\boldsymbol{\theta})$ by reparameterization, define a new set of parameters $\boldsymbol{\phi}$ for the model such that ϕ_1 is equal to $g(\boldsymbol{\theta})$:

$$\begin{aligned} \phi_1 &= \frac{\frac{995.0856}{1.632}\theta_1\theta_3}{1+\theta_2 2069+\theta_1 990.8+\theta_4 621.9} \\ \phi_2 &= \theta_2 \\ \phi_3 &= \theta_3 \\ \phi_4 &= \theta_4 \end{aligned} \tag{3.43}$$

Solving the set of equations in (3.43) for $\boldsymbol{\theta}$,

$$\begin{aligned} \theta_1 &= \frac{2.5\phi_1(1+2069\phi_2+621.9\phi_4)}{1524\phi_3-2477\phi_1} \\ \theta_2 &= \phi_2 \\ \theta_3 &= \phi_3 \\ \theta_4 &= \phi_4 \end{aligned} \tag{3.44}$$

The model is now written in terms of the new parameters:

$$\begin{aligned} y = &(\phi_1\phi_3(1+2069\phi_2+621.9\phi_4)(5x_2-3.06x_3)) \\ &\Big/ 2000\,(1.524\phi_3-2.477\phi_1+1.524\phi_2\phi_3x_1 \\ &-2.477\phi_1\phi_2x_1+0.0025\phi_1x_2+5.173\phi_1\phi_2x_2 \\ &+1.555\phi_1\phi_4x_2+1.524\phi_3\phi_4x_3-2.477\phi_1\phi_4x_3)+\epsilon \end{aligned} \tag{3.45}$$

The profile t plots for the $\boldsymbol{\phi}$ parameters are equal to the profile t plots generated

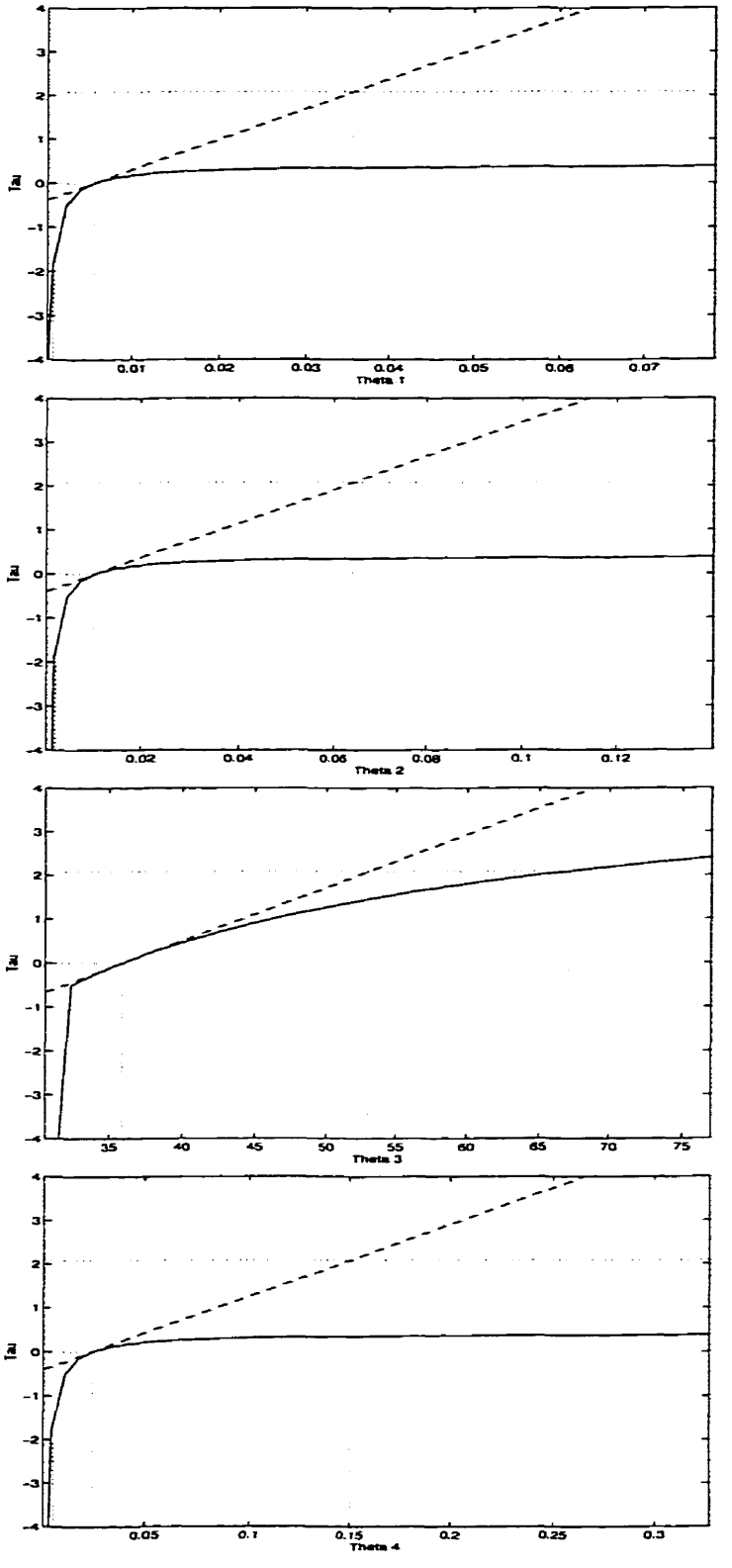

Figure 3.4: Profile t plots for the parameters of the isomerization model, generated using Chen's optimization algorithm (solid line: profile; dashed line: reference line; dotted lines: lines indicating critical values of τ and limits of inference intervals).

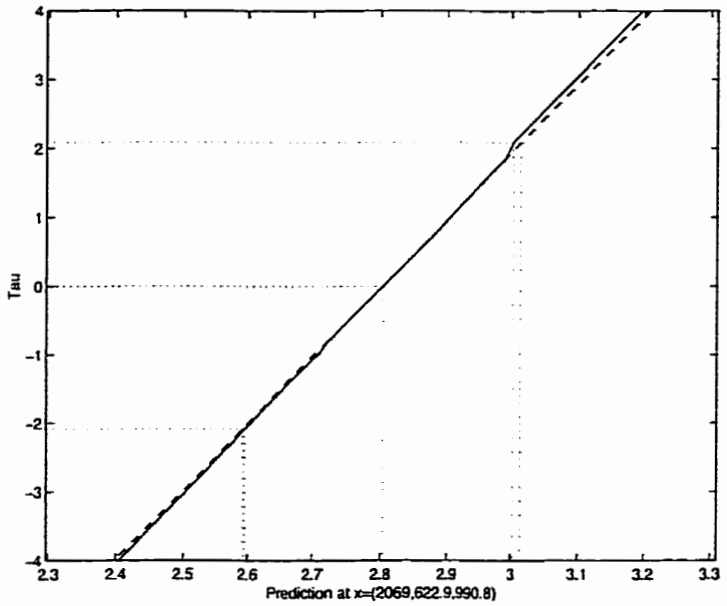

Figure 3.5: Profile t plot for the predicted reaction rate at $x = (2069, 990.8, 621.9)$ from the isomerization model, generated using the reparameterization algorithm (solid line: profile; dashed line: reference line; dotted lines: lines indicating critical values of τ and limits of inference intervals).

Table 3.2: Point Estimates and Inference Results for Example 1.

			Linearization		Profiling	
Statistic	MLE	se	L.B.	U.B.	L.B.	U.B.
θ_1	0.055	0.0145	-0.0248	0.0357	0.0007	∞
θ_2	0.0103	0.0259	-0.4380	0.0643	0.0019	∞
θ_3	35.9202	8.2123	18.7898	53.0506	26.4317	67.04
θ_4	0.0242	0.0603	-0.1016	0.1501	0.0049	∞
$y(2069, 990.8, 621.9)$	2.8032	0.1013	2.5918	3.0145	2.5946	3.0135
$y(734.98, 470.91, 72.39)$	6.5078	1.4056	4.8882	8.1275	5.1955	7.1737

by the optimization algorithm.

To obtain a $(1-\alpha)$ likelihood interval for $g(\boldsymbol{\theta})$ from the profile t plot, one must simply find the values of $g(\boldsymbol{\theta})$ on the profile t curve at which $\tau = \pm t(n-p;\alpha/2)$. For this example, the critical value of the t statistic is 2.086. Horizontal lines at ± 2.086 are drawn from the y-axis across to the profile t curve and then vertical lines are dropped down to the x-axis (see Figure 3.5). The values at which these vertical lines intersect the x-axis define the limits of the likelihood interval for $g(\boldsymbol{\theta})$. Note that all of the profile t plots shown in this work show values of τ ranging between ± 4. This is a relatively wide range for τ given that confidence levels of 95% and 99% are most commonly used and, depending on the number of degrees of freedom, those confidence levels correspond to τ values of approximately ± 2 and ± 3, respectively.

The value of $g(\boldsymbol{\theta})$ at $\tau = 0$ is the maximum likelihood estimate of $g(\boldsymbol{\theta})$. The profile t curve is tangent to the linear approximation reference line shown as a dashed line in these examples. Although it is easy to compute the limits of the linear approximation confidence interval using (3.12), these limits can also be read directly from the profile t plots. This is done in the same way as for the likelihood interval described above; however, for the linear approximation interval limits we consider values of $g(\boldsymbol{\theta})$ based on the dashed straight line. In this way, the profile t plots can be effective in illustrating the differences between the limits of a linear approximation confidence interval and the corresponding profiling likelihood interval.

From this example, it is important to appreciate that although the estimates of the parameters are drastically nonlinear, the prediction at $x = (2069, 990.8, 621.9)$ behaves relatively linearly over the region of interest, and that although the parameter estimates have likelihood intervals of infinite lengths, the uncertainty in the prediction is finite. Therefore, it is important to judge the value of a model based on inference results for the functions of interest, which may not necessarily be the parameters themselves.

The results for Example 1 can be rationalized in an intuitive and qualitative way. When a model is overparameterized with respect to the information contained in the set of data to which it was fit, only linear relationships between pairs of parameters can be effectively estimated, and so the uncertainties in the values of individual parameters may be high. However, the fitted model may still be able to describe the overall behavior of the data so that good predictions are possible. For example, it may be that there is not enough information in the data to estimate both θ_1 and θ_2 because these two parameters are approximately related as

$$\theta_1 \approx c\,\theta_2 \tag{3.46}$$

where c is a constant. In this case, the individual uncertainties in θ_1 and θ_2 will be high since any combination of θ_1 and θ_2 satisfying (3.46) will result in the same vector of predictions. However, the model expressed in terms of only θ_1 or only θ_2 may adequately represent the data and therefore good predictions may be obtained.

There are no strict rules by which to judge the relative linearity / nonlinearity of a function of parameters *a priori*. The prediction at $x = (2069, 990.8, 621.9)$ conforms with the claim of Clarke (1987) that model predictions tend to behave linearly even in cases where the parameters in the model show severe nonlinearity; however, a prediction at a different point for this same example shows why one should not place too much faith in that claim.

The profile t plot for the prediction at $x = (734.98, 470.91, 72.39)$ does not behave linearly (see Figure 3.6). This point is one of the eight extreme points of the space spanned by the experimental data. Although not shown here, profile t plots for the other seven extreme points were also generated. Only points (3247, 2030, 1083), (3247, 2030, 72.39) and (734.98, 470.91, 72.39) showed severe nonlinearity. The profile t plots for the remaining data points were similar to that for the prediction at $x =$

(2069, 990.8, 621.9) shown in Figure 3.5. Note that the estimates of the predictions at $x = (3247, 470.91, 1083)$ and $x = (734.98, 470.91, 1083)$ are negative. Neither the likelihood intervals nor the linearization intervals for these predictions included positive values. Clearly the model is inadequate for these combinations of pressures. To interpret the nonlinearity observed in the prediction at $x = (734.98, 470.91, 72.39)$,

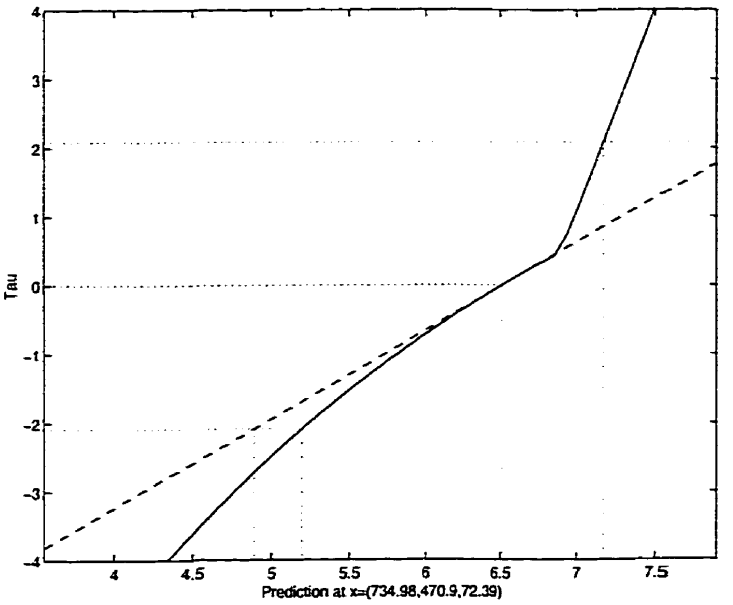

Figure 3.6: Profile t plot for the prediction at $x = (734.98, 470.91, 72.39)$ from the isomerization model generated using the reparameterization algorithm (solid line: profile; dashed line: reference line; dotted lines: lines indicating critical values of τ and limits of inference intervals).

it is helpful to look to the data and the model structure. From the plots of y versus each of the explanatory variables (Figure 3.7), it is clear that at low values of x_1, the measured rate of reaction y is high, and at low values of x_3, y takes on intermediate to high values. However, at low values of x_2, the measured reaction rate is low. For a prediction at low values of all three explanatory variables, there is a conflict among the effects of the three variables. The optimal prediction at this point is 5.91 which is a compromise between the high rates of reaction observed at low values of x_1 and x_3 and the low rates observed at low values of x_2. Perhaps because it was unrealistic to collect data at low concentrations of all three reactants, the data set

does not contain information at these concentrations. When a data set does not contain enough information to estimate a parameter or function of parameters, the likelihood intervals tend to become wide and this often manifests itself on profile t plots as a curve which is highly nonlinear. The nonlinearity is a result of the lack of information.

If more information were obtained by performing additional experiments, the estimates of the parameters could be improved, and their nonlinearity would likely be decreased. Ideally, if the interest really was in finding estimates of the parameters, further experimentation based on nonlinear design of experiments could be done. Bates and Watts (1988) used the isomerization example to discuss D-optimal designs. Profiling is a good way of identifying unacceptable levels of uncertainty in parameter estimates. Subset experimental designs can then be used to collect information about specific subsets of parameters which were identified as having been poorly estimated (Bates and Watts, 1988).

At this time, we are aware of no existing method for assessing the nonlinearity of a prediction prior to profiling. Although this example supports the conjecture that predictions at points well inside the experimental region tend to behave linearly, it also suggests that profiling is a prudent alternative to using linearization likelihood intervals.

3.8.2 Example 2

We use this example to illustrate a geometric interpretation of profiling inference results, and to explore further the use of profiling to obtain inference results for functions of the parameters of a fitted model. This example was taken from Draper and Smith (1981) (based on data reported by Smith and Dubey, 1964). Product A is produced with an initial fraction of available chlorine. Over time, this fraction of

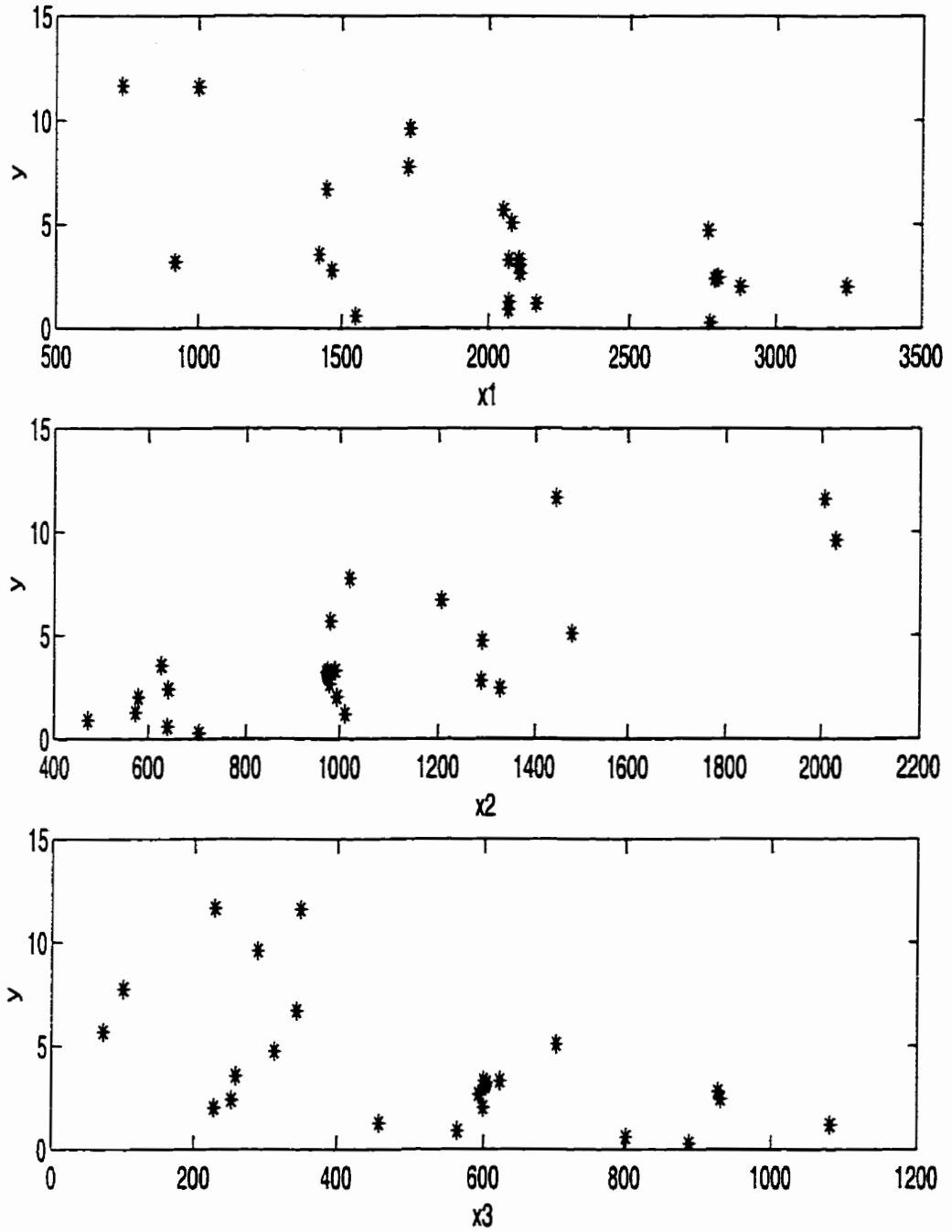

Figure 3.7: Plot of y versus each Explanatory Variable for Example 1.

available chlorine decreases according to the model:

$$y = \theta_1 + (0.49 - \theta_1)\exp(-\theta_2(t-8)) + \epsilon \tag{3.47}$$

where y is the fraction of available chlorine in Product A at time t. The data are listed in Table 3.3 and the maximum likelihood estimates of the parameters and associated inference results appear in Table 3.4. A plot of the data is shown in Figure 3.8. Profile t plots for the parameters are shown in Figure 3.9. The parameters behave relatively linearly as does the profile t plot for the predicted amount of available chlorine in Product A after 35 weeks (Figure 3.10). The linear approximation confidence intervals for these statistics therefore provide good approximations to the likelihood intervals.

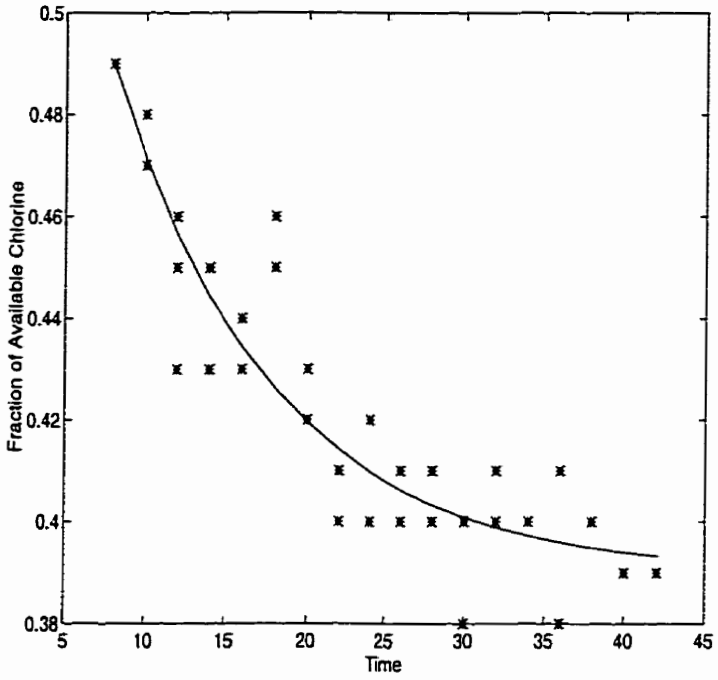

Figure 3.8: Plot of fraction of available chlorine versus time (in weeks) for Example 2.

However, even though linear approximation inference results are appropriate for the estimates of the parameters of a model, it does not follow that linear approximation will also provide reliable inference results for functions of the parameters. For the purpose of illustration, consider the prediction of the time at which the fraction

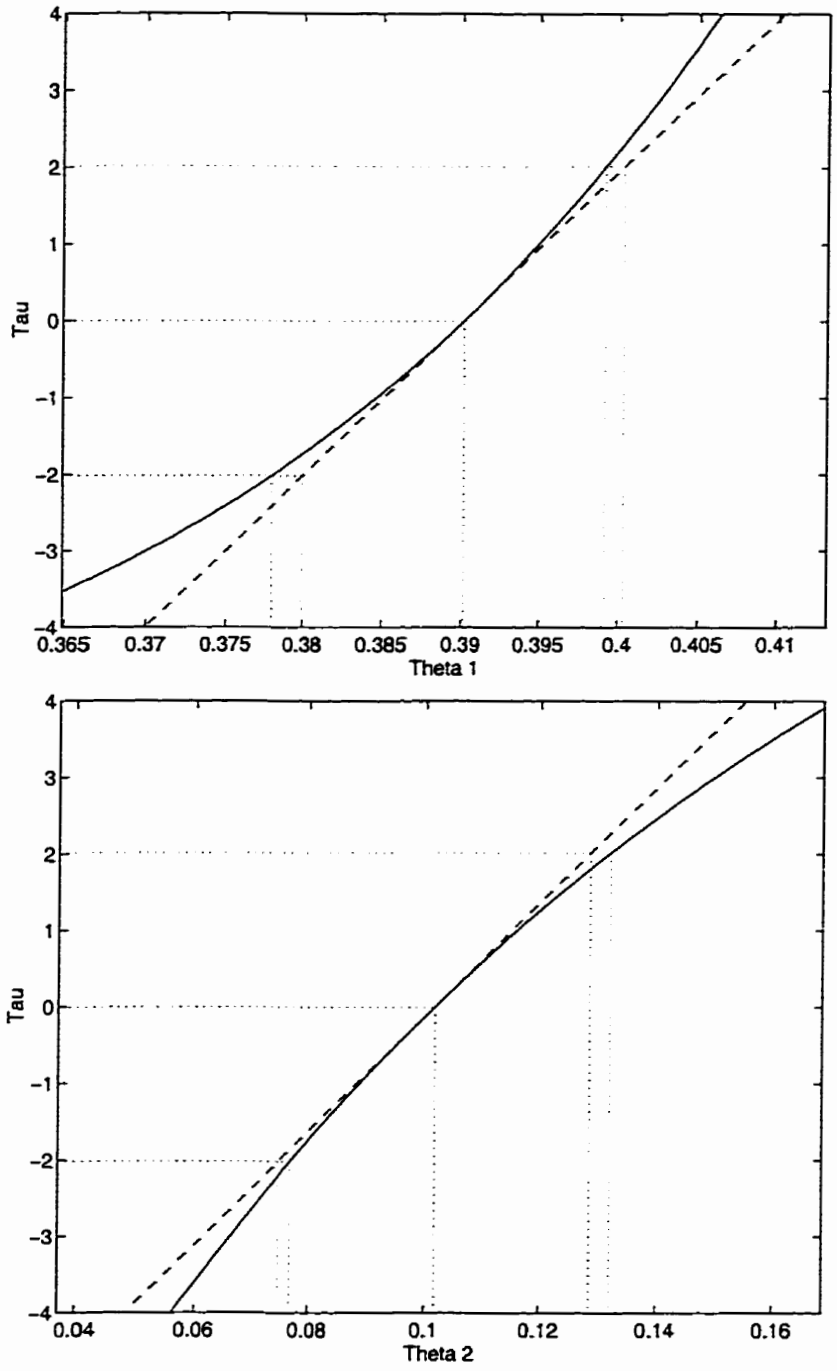

Figure 3.9: Profile t plots for the parameters of Example 2 (solid line: profile; dashed line: reference line; dotted lines: lines indicating critical values of τ and limits of inference intervals).

Table 3.3: Chlorine Data (Example 2).

Time t (weeks)	Fraction of Available Chlorine in Product A
8	0.49, 0.49
10	0.48, 0.47, 0.48, 0.47
12	0.46, 0.46, 0.45, 0.43
14	0.45, 0.43, 0.43
16	0.44, 0.43, 0.43
18	0.46, 0.45
20	0.42, 0.42, 0.43
22	0.41, 0.41, 0.40
24	0.42, 0.40, 0.40
26	0.41, 0.40, 0.40
28	0.41, 0.40
30	0.40, 0.40, 0.38
32	0.41, 0.40
34	0.40
36	0.41, 0.38
38	0.40, 0.40
40	0.39
42	0.39

Table 3.4: Point Estimates and Inference Results for Example 2.

			Profiling		Linearization	
Statistic	MLE	se	L.B.	U.B.	L.B.	U.B.
θ_1	0.3901	0.0050	0.3800	0.4003	0.3779	0.3992
θ_2	0.1016	0.0134	0.0747	0.1286	0.0768	0.1321
$y(t = 35weeks)$	0.3966	0.0029	0.3908	0.4024	0.391	0.402
$t(y = 0.40)$	30.7523	2.2396	26.0798	35.4820	27.388	45.2

of available chlorine remaining in Product A is 0.40. Then,

$$g(\boldsymbol{\theta}) = \frac{1}{\theta_2} ln \left(\frac{0.49 - \theta_1}{0.4 - \theta_1} \right) + 8 \tag{3.48}$$

The profile t plot for this function of the parameters is shown in Figure 3.11. It can be seen that $g(\boldsymbol{\theta})$ is very nonlinear, with an upper limit of approximately 45.2.

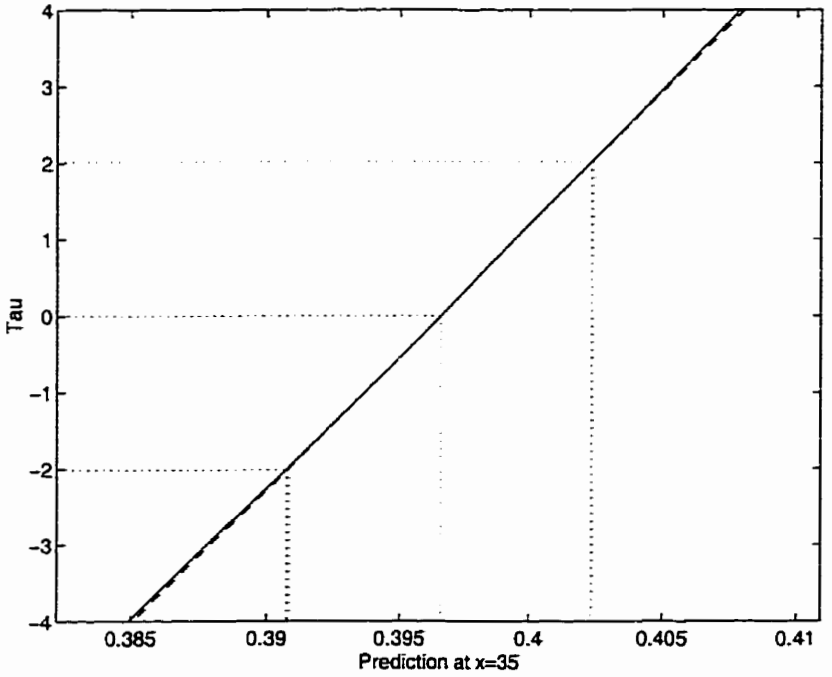

Figure 3.10: Profile t plot for the prediction at $t = 35$ for Example 2 (solid line: profile; dashed line: reference line; dotted lines: lines indicating critical values of τ and limits of inference intervals).

To interpret this result we again look to the data and the model structure. In this case, the parameters of the model have physical meaning. Parameter θ_1 represents the fraction of chlorine available in Product A a long time after the product was produced. Based on the structure of the model, the minimum fraction of available chlorine is approached asymptotically. Fitting the model to the data tells us that the maximum likelihood estimate of that minimum fraction of chlorine is 0.39. However, there is scatter in the data so this estimate is uncertain. The fact that the upper limit of the likelihood interval for the time at which the fraction of available chlorine=0.40

is very large reflects the statistical possibility that the minimum fraction of available chlorine may actually be 0.40 and therefore a large amount of time would have to elapse to achieve this level of available chlorine.

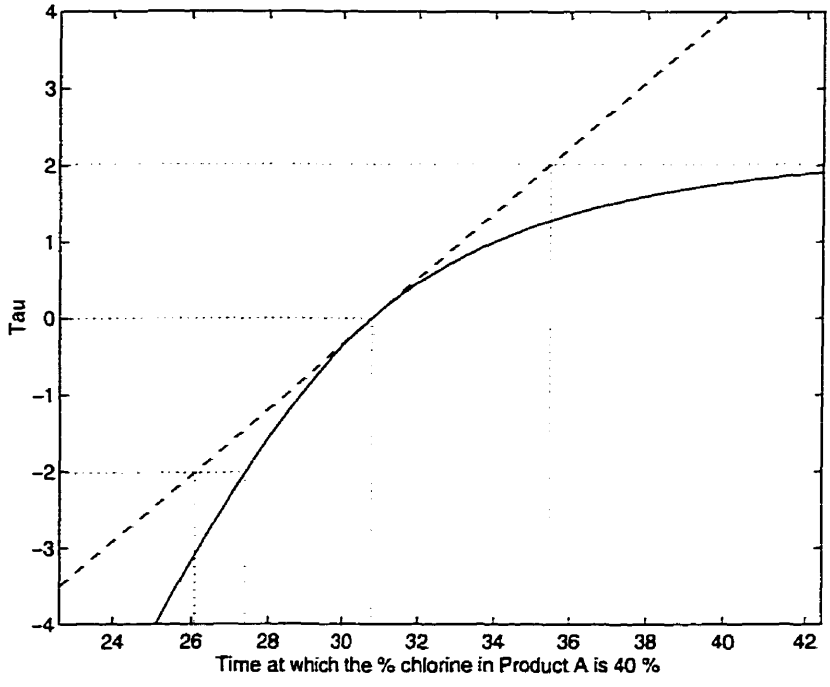

Figure 3.11: Profile t plot for the time at which the fraction of available chlorine =0.40 for Example 2 (solid line: profile; dashed line: reference line; dotted lines: lines indicating critical values of τ and limits of inference intervals).

Because the model for this example has only two parameters, it is possible to generate contour plots of the sum of squares surface and of the two functions of the parameters of interest. These contour plots are shown in Figures 3.12 and 3.13. The levels of the contours of the sums of squares surface are defined in terms of the nominal confidence level as determined by (3.39).

These plots help to illustrate the geometrical interpretation of profiling. The profiles identify the maximum and minimum values of $g(\boldsymbol{\theta})$ on or within the likelihood region for the parameters. In this case, $g(\boldsymbol{\theta})$ is a monotonic function of the parameters and the maximum and minimum values of a 95% likelihood interval for $g(\boldsymbol{\theta})$ occur at the points where the contours of $g(\boldsymbol{\theta})$ are tangent to the contour of $S(\boldsymbol{\theta})$ corresponding

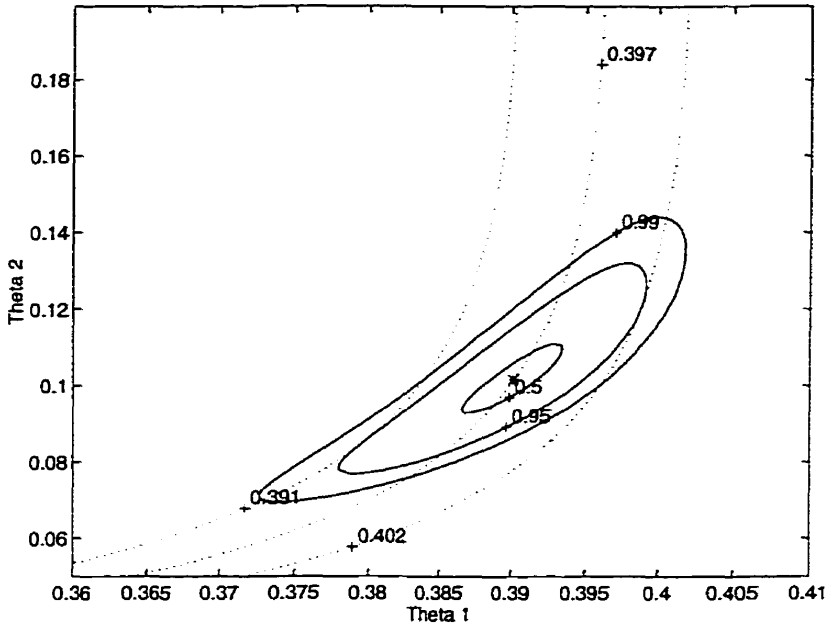

Figure 3.12: Contour plots of the sum of squares surface (solid lines) and the predicted fraction of available chlorine at $t = 35$ (dashed lines) for Example 2.

to the 95% confidence level. For this example neither the model nor $g(\boldsymbol{\theta})$ are linear and consequently the contours of $S(\boldsymbol{\theta})$ are not elliptical, as they would be for a model that is linear in the parameters. For the prediction of the time at which the fraction of available chlorine is 0.40, the contours of $g(\boldsymbol{\theta})$ become very closely spaced near the limit of its 95% likelihood interval, defined by the curves of $g(\boldsymbol{\theta})$ which are tangent to the likelihood region for the parameters. This representation reveals clearly why the upper limit of the 95% likelihood interval for $g(\boldsymbol{\theta})$ is nearly unbounded.

Empirically we have found that the nonlinearity of a function $g(\boldsymbol{\theta})$ will likely be high if its values approach an asymptotic limit, or its values are influenced by an asymptotic limit. This is the case for $g(\boldsymbol{\theta}) = \frac{1}{\theta_2} ln\left(\frac{0.49-\theta_1}{0.40-\theta_1}\right) + 8$ in Example 2. Nonlinearity is also to be expected if the model is being used to make predictions outside of the region of values within which the observations were made or in a region of values which is physically unrealistic. This is the case for the prediction of an extreme point discussed in Example 1.

In both examples we have assumed that the additive random error is indepen-

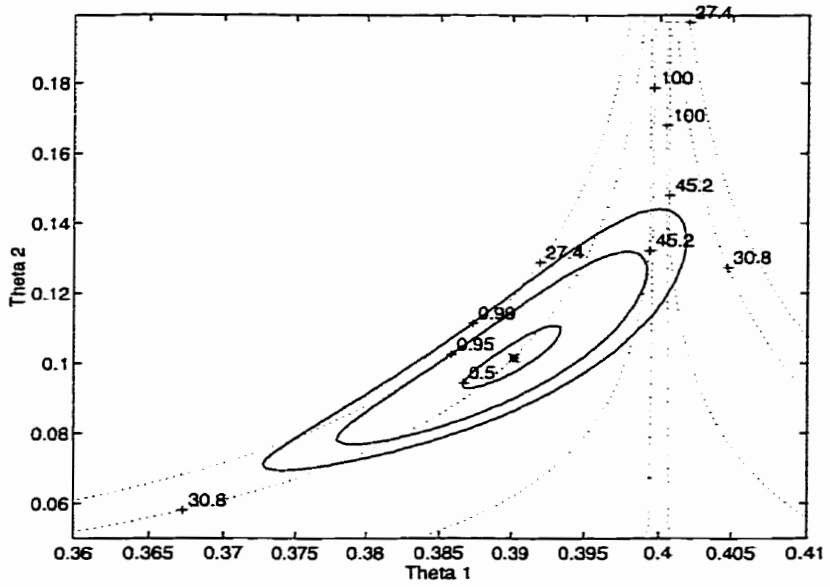

Figure 3.13: Contour plots of the sum of squares surface (solid lines) and the time at which the predicted fraction of available chlorine = 0.40 (dotted lines) for Example 2.

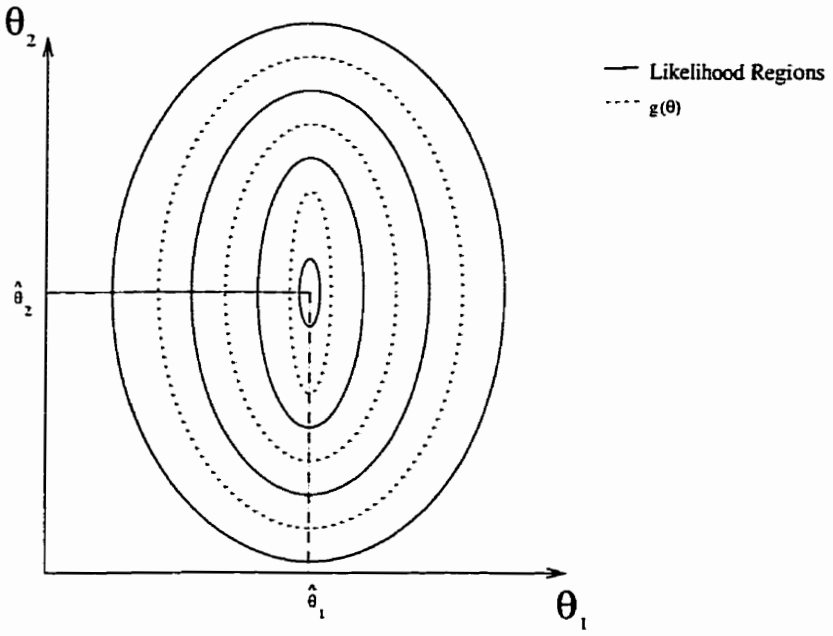

Figure 3.14: Contour plots of $g(\boldsymbol{\theta})$ and $L(\boldsymbol{\theta})$ for the case $g(\boldsymbol{\theta}) = kL(\boldsymbol{\theta})$.

dently and identically normally distributed. However, we emphasize that because profiling can be expressed in terms of a likelihood ratio, it is flexible enough to handle non-normal distributions. Alternatively, a transformation could be used to induce normality; for example, when the random error is proportional to the level of the response, a logarithmic transformation can be used to make the error appear iid normal.

3.9 Comments on computational issues

The choice of approach to generalized profiling is obviously at the discretion of the user. It has been our experience that neither the reparameterization approach nor the constrained optimization approach is consistently "better" than the other. With respect to computation time, the reparameterization approach often outperforms the constrained optimization approach because the unconstrained optimization of $(p-1)$ variables requires fewer function evaluations than the optimization of p variables subject to one constraint. For example, the CPU time required to construct the profile t plot for the prediction at $x = (2069, 990.8, 621.9)$ in Example 1 was 427.76 seconds using the optimization approach, whereas it was only 23.68 seconds using the reparameterization approach. These calculations were done using a SUN Ultra-1 workstation. It should be noted that a high-level programming language (MATLAB) was used, and no effort was made to make the code computationally effient. It is likely that the computation time could be reduced significantly by using an efficient code written in a low-level language. Nonetheless, both approaches typically perform well since good starting guesses for all of the optimization subproblems are almost always available. The starting guesses used in both algorithms are the same. The maximum likelihood estimates of the parameters are used as the starting guesses in the first iteration of the generalized profiling algorithm. These are likely to be good

guesses since, for any individual parameter or function of parameters, the algorithm proceeds away from the maximum likelihood estimate only in small steps. In subsequent iterations of the profiling algorithm, the location of the constrained optimum from the previous iteration is used as the starting guess.

The reparameterization approach to generalized profiling has been found to be preferable in cases in which the parameter effects nonlinearity is reduced by the reparameterization (Clarke, 1987). Such a reduction in nonlinearity can be quantified using measures of nonlinearity (e.g. Bates and Watts (1980)). However, it may be argued that the computational effort required to compute these measures of nonlinearity cannot be justified; it might be preferable to begin with the reparameterization approach to profiling from the outset. The reparameterization approach is advantageous only if the reparameterization can be done analytically. We have found that when numerical reparameterization of the model is required for a particular function $g(\boldsymbol{\theta})$, the constrained optimization approach consistently outperforms the reparameterization approach in terms of computation time.

Although the reparameterization approach may often prove to be the faster algorithm, we have found that the constrained optimization approach is easier to implement, and this may justify the added computational burden. The reparameterization approach involves finding analytical expressions for the new parameters in terms of the old parameters and $g(\boldsymbol{\theta})$, and subsequently expressing the model in terms of the new parameters. This can be readily accomplished using software for symbolic computation; however, in the absence of such software, the task can be time consuming and tedious. The reparameterization approach also appears to require a deeper understanding of the generalized profiling algorithm because appropriate manipulations of the model and the function of parameters is required before profiling can begin. This may inhibit its use.

Overall, the information collected throughout the profiling process is the same

regardless of the algorithm used. Bates and Watts (1988) have developed a means of sketching joint likelihood regions for parameters based on the results of profiling. Sketches of joint confidence regions for parameters and functions of parameters can be created from the results of either of the two approaches discussed in this paper.

3.10 When profiling fails

There exists a special class of functions $g(\boldsymbol{\theta})$ for which profiling is not an appropriate method for computing likelihood intervals. This class was discussed in Quinn et al. (1999b) (Chapter 4) and is examined here briefly.

When an unconstrained optimum of $g(\boldsymbol{\theta})$ is located at the same point as the unconstrained maximum of $L(\boldsymbol{\theta})$, the profiling algorithm will fail. When the least squares estimate of $\boldsymbol{\theta}$ is also the location of the unconstrained optimum of $g(\boldsymbol{\theta})$, then

$$\frac{\partial g}{\partial \boldsymbol{\theta}}\Big|_{\hat{\boldsymbol{\theta}}} = \hat{\dot{g}} = 0 \tag{3.49}$$

and

$$\begin{aligned} se(\hat{g}) &= \hat{\dot{g}}' \left(\frac{\partial f}{\partial \boldsymbol{\theta}}' \frac{\partial f}{\partial \boldsymbol{\theta}} \right)^{-1} \hat{\dot{g}} \\ &= 0 \end{aligned} \tag{3.50}$$

where $se(\hat{g})$ is the standard error of $g(\hat{\boldsymbol{\theta}})$. The profiling algorithm fails since its step sizes are based on multiples of $se(\hat{g})$. Furthermore, the reference line can not be plotted since it is defined by

$$\tau_{lin} = \frac{g_c - \hat{g}}{se_g} \tag{3.51}$$

These problems associated with the profiling algorithm do *not* imply that a likeli-

hood interval for $g(\theta)$ cannot be found. They simply suggest that an alternate method is required.

It is instructive to consider a particular case in more detail. Consider

$$g(\boldsymbol{\theta}) = kL(\boldsymbol{\theta}) \tag{3.52}$$

where k is a scalar and $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$.

In this case the contours of $g(\boldsymbol{\theta})$ will be coincident with the contours of $L(\boldsymbol{\theta})$, although the levels associated with the contours of the two functions will be different. If $L(\boldsymbol{\theta})$ is a quadratic function, then the contours of $g(\boldsymbol{\theta})$ and $L(\boldsymbol{\theta})$ will be a series of concentric ellipses. This is illustrated in Figure 3.14 for the case of normally distributed random errors, where maximizing $L(\boldsymbol{\theta})$ is equivalent to minimizing $S(\boldsymbol{\theta})$. Note that Figure 3.14 illustrates a case where the parameters are uncorrelated, but this need not be so in general.

The upper limit of a likelihood interval for $g(\boldsymbol{\theta})$ is defined to be the maximum value of $g(\boldsymbol{\theta})$ over the likelihood region for $\boldsymbol{\theta}$, which is defined by the function $S(\boldsymbol{\theta})$. The lower limit is the minimum of g($\boldsymbol{\theta}$) over the likelihood region. Consequently, one limit of the likelihood interval for $g(\boldsymbol{\theta})$ is equal to the value of the contour of $g(\boldsymbol{\theta})$ which is coincident with the contour for $S(\boldsymbol{\theta})$ having the critical value S_{crit}, where S_{crit} is that value of $S(\boldsymbol{\theta})$ which satisfies

$$t(n-p, \alpha/2) = \sqrt{\frac{S_{crit} - S(\hat{\theta})}{s^2}} \tag{3.53}$$

These ideas can be stated in the form of an optimization problem as follows:

$$\begin{aligned} &\text{Optimize} && g(\boldsymbol{\theta}) \\ &\text{Subject to} && S(\boldsymbol{\theta}) \leq S_{crit} \end{aligned} \tag{3.54}$$

For this case the location of one optimum of $g(\boldsymbol{\theta})$ is the locus of points defining the ellipsoid $S(\boldsymbol{\theta}) = S_{crit}$, or equivalently, $g(\boldsymbol{\theta}) = kL_{crit}$.

Note that this limit of the likelihood interval for $g(\boldsymbol{\theta})$ is uniquely defined despite the fact that the optimum of $g(\boldsymbol{\theta})$ occurs at an infinite number of points. That is, although the location of the optimum of $g(\boldsymbol{\theta})$ is not unique, the optimum itself is, and so the limit for the likelihood interval for $g(\boldsymbol{\theta})$ is also uniquely defined. When only the likelihood interval for $g(\boldsymbol{\theta})$ is of interest, the non-uniqueness of the location of the optimum is of little consequence. However, determining joint likelihood regions for $g_i(\boldsymbol{\theta})$ and $g_j(\boldsymbol{\theta})$, or for $g_i(\boldsymbol{\theta})$ and θ_q, where $g_i(\boldsymbol{\theta})$ and $g_j(\boldsymbol{\theta})$ are any two functions of parameters of interest, would be a problem.

The other limit of the likelihood interval of $g(\boldsymbol{\theta})$ is $g(\hat{\boldsymbol{\theta}})$; it is independent of α and requires no optimization. Measures of system performance commonly used in engineering are examples of functions $g(\boldsymbol{\theta})$ which have their minima at $\hat{\boldsymbol{\theta}}$. For example, the measure of controller performance $\Delta Perf$ used by Shirt et al. (1994) is one example of a $g(\boldsymbol{\theta})$ which can have a minimum at $\hat{\boldsymbol{\theta}}$.

3.11 Conclusion

Bates and Watts (1988) proposed a profiling algorithm as a means of finding reliable likelihood intervals for parameters in nonlinear models. Theoretical development of the generalization of this algorithm for functions of parameters has been presented in detail in this paper. The generalized profiling algorithm is appropriate for finding likelihood intervals for functions of parameters for several classes of models, so long as the likelihood ratios for those functions of parameters can be determined. Two different approaches to the generalization have been discussed, and both approaches have been used to elucidate the merits and limitations of the generalized algorithm.

The profiling algorithm provides likelihood intervals which account for the non-

linearity of a model and the profile t plots provide qualitative information about the nonlinearity of the solution surface. This is helpful in judging the behaviour of the estimates of interest.

There is much to recommend the profiling algorithm; however, the algorithm does fail when the location of the unconstrained minimum of $g(\boldsymbol{\theta})$ is located at $\hat{\boldsymbol{\theta}}$. In this case, $\hat{\boldsymbol{\theta}}$ is the location of the lower limit of the likelihood interval. The upper limit can be found by maximizing $g(\boldsymbol{\theta})$ subject to $L(\boldsymbol{\theta}) \geq L_{crit}$.

Both the reparameterization and the constrained optimization approaches to generalizing the profiling algorithm are helpful in understanding the algorithm. However, several related issues remain to be investigated further. These include: sketching joint confidence regions, measuring coverage probabilities and applying measures of nonlinearity. Some work has been done in the area of diagnostics of nonlinearity (Linssen, 1975; Bates and Watts, 1980; Cook and Goldberg, 1986, among others). Use of such diagnostic tools may alleviate the need for profiling in cases where nonlinearity is low; however, it may be argued that in many cases the effort required to compute the diagnostics is comparable to that required to compute the profile t plots.

3.12 Acknowledgments

The authors gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada and the School of Graduate Studies of Queen's University. We thank the referees for their thorough review. Their constructive feedback has helped to improve this paper.

3.13 Nomenclature

$\boldsymbol{a}$	$= p \times 1$ vector of constants
c	= a constant
$D^2(\boldsymbol{\theta})$	= deviance
$\boldsymbol{e}$	$= n \times 1$ column vector of estimated random errors
$F(p, n-p; \alpha)$	= upper α quantile for the F distribution with p and $n-p$ degrees of freedom
$g(\boldsymbol{\theta})$	a function of parameters
$L(\boldsymbol{\theta})$	= likelihood function evaluated at $\boldsymbol{\theta}$
LR	= likelihood region
m	= number of regressor variables
n	= number of observations
p	= number of estimated parameters
s	= estimated standard deviation of the random errors
$S(\boldsymbol{\theta})$	= sum of squared errors
se	= standard error
t	= time
$t(n-p; \alpha/2)$	= upper $\alpha/2$ quantile of the t distribution with $n-p$ degrees of freedom
t_{max}	= time at which the concentration of Species B reaches a maximum (in weeks)
$\boldsymbol{V}$	$= n \times p$ matrix of elements v_{ij} representing the first derivative of $f(\boldsymbol{x}_i, \boldsymbol{\theta})$ with respect to the j^{th} parameter
$\boldsymbol{x}$	$= 1 \times m$ row vector of m independent variables
$\boldsymbol{X}$	$= n \times p$ matrix of elements x_{ij} representing the level of the j^{th} independent variable for observation i

y = response variable

$\boldsymbol{y}$ = $n \times 1$ column vector of values of the response variable

$z^*(c)$ = signed root deviance

Greek letters

α = significance level

$\delta(g(\boldsymbol{\theta}))$ = studentized value of $g(\boldsymbol{\theta})$

ϵ = additive random error

$\boldsymbol{\epsilon}$ = $n \times 1$ column vector of random errors

θ_q = q^{th} parameter of a model

$\boldsymbol{\theta}$ = $p \times 1$ column vector of parameters of a model

σ = standard deviation

$\tau(g(\boldsymbol{\theta}))$ = profile t statistic for $g(\boldsymbol{\theta})$

ϕ_q = q^{th} parameter of a reparameterized model

$\boldsymbol{\phi}$ = $p \times 1$ column vector of an alternate set of parameters of a model

$\chi^2(1)$ = the chi-squared distribution with 1 degree of freedom

Superscripts

$*$ = a true value

$\hat{}$ = a maximum likelihood estimate

$\tilde{}$ = a constrained estimate

Abbreviations

iid	independently and identically distributed
L.B.	lower bound
U.B.	upper bound
MLE	maximum likelihood estimate

Chapter 4

A Note on Likelihood Intervals and Profiling

4.1 Abstract

In many applications, decisions are made on the basis of a function of parameters $g(\boldsymbol{\theta})$. When the value of $g(\boldsymbol{\theta})$ is calculated using estimated values for the parameters, it is important to have a measure of the uncertainty associated with that value of $g(\boldsymbol{\theta})$. Likelihood ratio approaches to finding likelihood intervals for functions of parameters have been shown to be more reliable, in terms of coverage probability, than the linearization approach. Two approaches to the generalization of the profiling algorithm have been proposed in the literature to enable construction of likelihood intervals for a function of parameters (Chen and Jennrich, 1996; Bates and Watts, 1988). In this paper we show the equivalence of these two methods. We also provide an analysis of cases in which neither profiling algorithm is appropriate. For one of these cases an alternate approach is suggested. Whereas generalized profiling is based on maximizing the likelihood function given a constraint on the value of $g(\boldsymbol{\theta})$, the alternative algorithm is based on optimizing $g(\boldsymbol{\theta})$ over a likelihood region.

4.2 Introduction

The use of likelihood ratios to make inference statements about parameters and functions of parameters in proposed models has a long history that can be traced back to Fisher (1939). However, it was not until the early 1980's that the benefits of this approach, relative to the computationally simpler linearization approach, were highlighted. Profile t plots and profiling were discussed by Bates and Watts (1988) for the special case of making inferences about parameters of nonlinear regression models. The idea of extending profiling-type algorithms to deal with inferences about functions of parameters has been proposed often in the literature (Chen and Jennrich, 1996; Chen, 1991; Ross, 1990; Bates and Watts, 1988; Clarke, 1987; Ratkowsky, 1983). However, in many cases, the theory was neither developed nor used. Chen (1991) provided a comprehensive explanation of his approach; however his conceptualization of the problem differed from that of the others. Whereas Chen approached the problem from an optimization perspective, most others perceived the generalization to be a reparameterization problem.

In this paper, both approaches are examined and their equivalence is shown. Some limitations of the profiling algorithm are discussed in light of the insights gained through the consolidation of the theory.

4.3 Background on Profiling

When developed in terms of likelihood ratios, the profiling algorithm is applicable to several classes of models. Here we consider a general model for a single response variable:

$$y = f(\boldsymbol{x}, \boldsymbol{\theta}) + \epsilon \tag{4.1}$$

where the function $f(\boldsymbol{x}, \boldsymbol{\theta})$ is the expected value of the response variable y at specified levels of m independent variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)^T$, and specified values of p parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$, and ϵ is an additive random error term associated with y. We assume everywhere that the random error is independently and identically distributed, but we make no assumptions about the distribution of ϵ. Note that in the case of discrete dynamic models, y may represent an observed value in a time series and $\boldsymbol{x}$ may represent a vector of variables which could include lagged values of the response y, and past and present values of other regressor (i.e., input) variables. The vector $\boldsymbol{x}$ may also include lagged values of ϵ so as to account for serial correlation in the data.

Profiling is based on testing the hypothesis that $g(\boldsymbol{\theta}) = c$ versus the alternative that $g(\boldsymbol{\theta}) \neq c$, where $g(\boldsymbol{\theta})$ is a function of the parameters in a proposed model, and c is a constant. The likelihood ratio for testing the null hypothesis is:

$$LR(c) = \frac{L(\tilde{\boldsymbol{\theta}})}{L(\hat{\boldsymbol{\theta}})} \tag{4.2}$$

where $L(\hat{\boldsymbol{\theta}})$ is the likelihood function evaluated at $\hat{\boldsymbol{\theta}}$, the unconditional maximum likelihood estimates of all parameters in the model, $L(\tilde{\boldsymbol{\theta}})$ is the conditional likelihood given that $g(\tilde{\boldsymbol{\theta}}) = c$, and $\tilde{\boldsymbol{\theta}}$ is the location of the conditional maximum. The expression for $L(\boldsymbol{\theta})$ depends on the assumptions about ϵ. It can be shown that, under the null hypothesis, the asymptotic distribution of $-2\ln(LR)$ is $\chi^2(1)$, where $\chi^2(1)$ is the chi-squared distribution with one degree of freedom (Lindgren, 1976). When the random errors ϵ are iid normal, $f(\boldsymbol{x}, \boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$ and the variance of the random error is known, then $-2\ln(LR)$ follows exactly the $\chi^2(1)$ distribution. In generalized profiling it is assumed that only an estimate of the variance of the random error is available, and therefore the limits of the likelihood intervals are based on critical values of $F(1, n-p; \alpha)$ so as to account for the uncertainty in the estimate of σ^2.

Chen and Jennrich (1996) defined the statistics:

$$D^2(\boldsymbol{\theta}) = -2\ln\left(\frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})}\right) \tag{4.3}$$

and

$$z^*(c) = sign(c - g(\hat{\boldsymbol{\theta}}))D(\bar{\boldsymbol{\theta}}) \tag{4.4}$$

where $D^2(\boldsymbol{\theta})$ is called the deviance, and $z^*(c)$ is called the signed root deviance (SRD) which is used as the basis for a profile plot called the signed root deviance profile (SRDP). Barndorf-Nielson (1986) defined a statistic similar to the SRD. The SRD is a generalization of the τ statistic of Bates and Watts (1988) and the SRDP is a generalization of their profile t plot. When the random errors are iid normal and critical values of the student t distribution are used to account for uncertainty in the estimate of the pure error variance, the SRDP and the profile t plot are equivalent. In this paper, the likelihood ratio plots will be referred to as profile t plots to be consistent with the early work. The term *generalized profiling* will be used to emphasize the use of profile plots in a broader sense than originally proposed, for predicted values of the response variable or functions of the parameters, for example. The emphasis in this paper is on the use of profiling for an arbitrary function of the parameters of a proposed model, noting that profiling an individual parameter θ_q is the special case where $g(\boldsymbol{\theta}) = \theta_q$.

4.4 The Optimization Approach

Chen (1991) developed a method for obtaining profiling results for an arbitrary function $g(\boldsymbol{\theta})$ by posing the problem in terms of a constrained optimization:

Maximize

$$L(\boldsymbol{\theta}) \tag{4.5}$$

subject to the constraint

$$g(\boldsymbol{\theta}) = c$$

To profile $g(\boldsymbol{\theta})$ is to solve a series of these optimization problems for a range of values of c above and below the maximum likelihood estimate of $g(\boldsymbol{\theta})$. We assume that $L(\boldsymbol{\theta})$ and $g(\boldsymbol{\theta})$ are twice differentiable with respect to $\boldsymbol{\theta}$.

For a model with p parameters fitted to a set of n data, the commonly used $100(1-\alpha)$ percent linearization interval for g (assuming $\epsilon \sim$ *iid* $N(0, \sigma^2)$) is:

$$g(\boldsymbol{\theta}) = g(\hat{\boldsymbol{\theta}}) \pm se(\hat{g})\, t(n-p; \alpha/2) \tag{4.6}$$

where

$$\{se(\hat{g})\}^2 = s^2 \left.\frac{dg}{d\boldsymbol{\theta}}^T\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{V}^T\boldsymbol{V})^{-1} \left.\frac{dg}{d\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{4.7}$$

and $\boldsymbol{V}$ is the $n \times p$ matrix with elements defined by $\boldsymbol{V}_{ij} = \left.\frac{\partial f(x_i,\theta)}{\partial \theta_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$, and $s^2 = \sum_{i=1}^{n} \left\{y_i - f(\hat{\boldsymbol{\theta}})\right\}^2 / (n-p)$ is an estimate of σ^2 with $(n-p)$ degrees of freedom (Chen, 1991).

4.5 An Equivalent Alternative

Finding likelihood intervals for functions of parameters has most often been posed as a reparameterization problem(Ross, 1990; Bates and Watts, 1988; Clarke, 1987; Ratkowsky, 1983). In this approach, a new set of parameters $\boldsymbol{\phi}$ is defined for the model such that $g(\boldsymbol{\theta})$ is one of the new parameters. Then, the algorithm for profiling parameters can be used directly. For example, the new parameters may be defined as:

$$\begin{aligned} \phi_1 &= g(\boldsymbol{\theta}) \\ \phi_2 &= \theta_2 \\ &\vdots \\ \phi_p &= \theta_p \end{aligned} \tag{4.8}$$

where $g(\boldsymbol{\theta})$ is the function of interest.

The reparameterization proceeds by solving the set of equations given in (4.8) for $\boldsymbol{\theta}$, such that

$$\begin{aligned} \theta_1 &= g^I(\boldsymbol{\phi}) \\ \theta_2 &= \phi_2 \\ &\vdots \\ \theta_p &= \phi_p \end{aligned} \tag{4.9}$$

where g^I is the first component of the inverse of the reparameterization defined by (4.8). Necessary assumptions about this inverse are discussed in Section 4.6. These results are then substituted into (4.1), and the model becomes

$$y = f(\boldsymbol{x}, \boldsymbol{\phi}) + \boldsymbol{\epsilon} \tag{4.10}$$

Profile t plots are based on the τ statistic (or the $z^*(c)$ statistic), and the quantity which defines this statistic is the conditional maximum of the likelihood function, $L(\tilde{\phi}) = \tilde{L}$. The fundamental difference between the reparameterization approach and the optimization approach is the way in which $\tilde{L}$ is computed.

Note that a maximization problem must be solved in both approaches. In Chen's approach, the optimization problem can be expressed as in (4.5). In the reparameterization approach, the optimization problem can be expressed as:

Maximize

$$L(\boldsymbol{\phi}) \tag{4.11}$$

subject to the constraint

$$\phi_1 = c$$

where $L(\boldsymbol{\phi})$ is the likelihood function for the reparameterized model.

To show the equivalence of the two approaches, the equivalence of the solutions to the two optimization problems is now demonstrated. Using the method of Lagrange multipliers (Edgar and Himmelblau, 1988), the solution to Chen's optimization problem is the solution to the set of equations:

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{\theta}} + \lambda_1 \frac{\partial g}{\partial \boldsymbol{\theta}} &= 0 \\ g(\boldsymbol{\theta}) - c &= 0 \end{aligned} \tag{4.12}$$

Similarly, the solution to the reparameterization problem is the solution to the set of

equations:

$$\begin{aligned} \frac{\partial L}{\partial \phi_1} + \lambda_2 &= 0 \\ \frac{\partial L}{\partial \phi_2} &= 0 \\ &\vdots \\ \frac{\partial L}{\partial \phi_p} &= 0 \\ \phi_1 - c = g(\theta) - c &= 0 \end{aligned} \tag{4.13}$$

By the chain rule,

$$\frac{\partial L}{\partial \boldsymbol{\phi}} = \left(\frac{\partial L}{\partial \boldsymbol{\theta}}\right)^T \left[\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}\right]^{-1} \tag{4.14}$$

Also, it can be shown that

$$\left[\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}\right]^{-1} = \tag{4.15}$$

$$\begin{bmatrix} [\frac{\partial g}{\partial \theta_1}]^{-1} & -\frac{\partial g}{\partial \theta_2}[\frac{\partial g}{\partial \theta_1}]^{-1} & \cdots & -\frac{\partial g}{\partial \theta_p}[\frac{\partial g}{\partial \theta_1}]^{-1} \\ 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Substituting (4.16) into (4.14),

$$\frac{\partial L}{\partial \boldsymbol{\phi}} = \begin{bmatrix} \frac{\partial L}{\partial \theta_1}\left[\frac{\partial g}{\partial \theta_1}\right]^{-1} \\ -\frac{\partial L}{\partial \theta_1}\frac{\partial g}{\partial \theta_2}\left[\frac{\partial g}{\partial \theta_1}\right]^{-1} + \frac{\partial L}{\partial \theta_2} \\ \vdots \\ -\frac{\partial L}{\partial \theta_1}\frac{\partial g}{\partial \theta_p}\left[\frac{\partial g}{\partial \theta_1}\right]^{-1} + \frac{\partial L}{\partial \theta_p} \end{bmatrix} \tag{4.16}$$

Therefore the set of equations given in (4.13) becomes

$$\begin{aligned} \frac{\partial L}{\partial \theta_1}\left[\frac{\partial g}{\partial \theta_1}\right]^{-1} + \lambda_2 &= 0 \\ -\frac{\partial L}{\partial \theta_1}\frac{\partial g}{\partial \theta_2}\left[\frac{\partial g}{\partial \theta_1}\right]^{-1} + \frac{\partial L}{\partial \theta_2} &= 0 \\ &\vdots \\ -\frac{\partial L}{\partial \theta_1}\frac{\partial g}{\partial \theta_p}\left[\frac{\partial g}{\partial \theta_1}\right]^{-1} + \frac{\partial L}{\partial \theta_p} &= 0 \\ g(\theta) - c &= 0 \end{aligned} \tag{4.17}$$

For the two approaches to be equivalent, the systems of equations given in (4.12) and (4.17) must be equal. Multiplying the first equation of set (4.17) by $\frac{\partial g}{\partial \theta_1}$,

$$\frac{\partial L}{\partial \theta_1} + \lambda_2 \frac{\partial g}{\partial \theta_1} = 0 \tag{4.18}$$

which is equal to the first equation of set (4.12). Also, rearranging (4.18),

$$-\frac{\partial L}{\partial \theta_1}\left[\frac{\partial g}{\partial \theta_1}\right]^{-1} = \lambda_2 \tag{4.19}$$

Using (4.19) to express the set of equations (4.17) in terms of λ_2,

$$\frac{\partial L}{\partial \boldsymbol{\theta}} + \lambda_2 \frac{\partial g}{\partial \boldsymbol{\theta}} = 0 \tag{4.20}$$

which is equivalent to the corresponding set of equations in (4.12) with $\lambda_1 = \lambda_2$. Thus, it has been shown that any solution to the constrained optimization problem in (4.5) is also a solution to the reparameterization problem in (4.11). However, the Lagrange method employed for the proof represents only a necessary condition for an extremum. Therefore, a solution to (4.11) could be a local minimum or saddle point. When the reparameterization approach is employed, care should be taken to ensure that the optimum is in fact a maximum. In practice, the solutions to the reparameterization problem and the constrained optimization problem are solved numerically, and there is risk that a local maximum, rather than the global maximum, will be found. See Mangasarian (1994) for assumptions necessary to ensure that the global maximum is found.

Chen (1991) asserted that for functions $g(\boldsymbol{\theta})$, the slope of the profile-t plot at $g(\hat{\boldsymbol{\theta}})$ is not necessarily equal to the slope of the linear approximation result. This is not consistent with the fact that Chen's approach to profiling is equivalent to the reparameterization approach. By the reparameterization method, the slope of the linear approximation is necessarily equal to the slope of the profile t plot at the least squares estimate, since it has been shown by Watts (1993) (in Agrawal, 1993) that for the case of a linear function of the parameters of a linear model, the $\tau(\boldsymbol{\theta})$ statistic is equal to

$$\delta(\boldsymbol{\theta}) = \frac{g(\boldsymbol{\theta}) - g(\hat{\boldsymbol{\theta}})}{se(g(\hat{\boldsymbol{\theta}}))} \tag{4.21}$$

which is the basis for the linearization intervals and the reference line on the profile t plot. The basis for Chen's assertion was the result that the slope of the profile t function for $g(\boldsymbol{\theta})$ at $\tau = 0$ is given by:

$$\left.\frac{\partial \tau}{\partial g}\right|_{g=\hat{g}} = \left(\sqrt{s^2 \frac{\partial g}{\partial \boldsymbol{\theta}}^T \left(\frac{1}{2} \frac{\partial^2 S}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} \right)^{-1} \frac{\partial g}{\partial \boldsymbol{\theta}}} \Big| \hat{\boldsymbol{\theta}} \right)^{-1} \tag{4.22}$$

where $S(\boldsymbol{\theta}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the sum of squared residuals, but the slope of the linear approximation profile plot is:

$$\left(\sqrt{s^2 \frac{\partial g}{\partial \boldsymbol{\theta}}^T (\boldsymbol{V}^T \boldsymbol{V})^{-1} \frac{\partial g}{\partial \boldsymbol{\theta}}\Big|_{\hat{\boldsymbol{\theta}}}} \right)^{-1} \tag{4.23}$$

However, Chen and Jennrich (1996) later showed that for the more general case where the additive noise is not assumed to be normally distributed (i.e where the results are derived in terms of the likelihood function $L(\boldsymbol{\theta})$), the slope of the SRDP at $g(\hat{\boldsymbol{\theta}})$ is equal to $1/se(\hat{g})$, where $se(\hat{g})$ is computed from the observed information matrix. The seemingly contradictory results of Chen (1991) and Chen and Jennrich (1996) stem from the assumptions made about the estimation of $se(\hat{g})$.

For large samples, the variance-covariance matrix for $\hat{\boldsymbol{\theta}}$, denoted by $\boldsymbol{\Omega}_{\hat{\boldsymbol{\theta}}}$, is asymptotically equal to $\boldsymbol{\Omega}_{\boldsymbol{\theta}_0}$, where

$$\begin{aligned} \boldsymbol{\Omega}_{\boldsymbol{\theta}_0} &= E\left\{ \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ln L(\boldsymbol{\theta})\right] \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ln L(\boldsymbol{\theta})\right]^T \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right\} \\ &= -E\left\{ \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ln L(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right\} \end{aligned} \tag{4.24}$$

where $\boldsymbol{\theta}_0$ denotes the true value of $\boldsymbol{\theta}$ (Ljung, 1987). For the specific case of independently and identically normally distributed error, it can readily be shown that

$$\begin{aligned} \boldsymbol{\Omega}_{\boldsymbol{\theta}_0} &= E\left\{ -\frac{1}{2\sigma^2} \frac{\partial^2}{\partial \boldsymbol{\theta}^2} S(\boldsymbol{\theta}) \right\}^{-1} \\ &= \sigma^2 \left(\frac{\partial f}{\partial \boldsymbol{\theta}}^T \frac{\partial f}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right)^{-1} \end{aligned} \tag{4.25}$$

so that for large samples, (4.22) and (4.23) are equivalent and are both equal to $1/se(\hat{g})$. However, in practice, the small sample estimate of the standard error of g may be poor, and the value of the estimate may depend on whether it is computed

using:

$$\Omega_{\hat{\boldsymbol{\theta}}} \approx \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ln L(\boldsymbol{\theta})\right]\left[\frac{\partial}{\partial \boldsymbol{\theta}} \ln L(\boldsymbol{\theta})\right]^T\Big|_{\hat{\boldsymbol{\theta}}} \tag{4.26}$$

or

$$\Omega_{\hat{\boldsymbol{\theta}}} \approx \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ln L(\boldsymbol{\theta})\Big|_{\hat{\boldsymbol{\theta}}} \tag{4.27}$$

which in the case of iid normal additive error are:

$$\Omega_{\hat{\boldsymbol{\theta}}} \approx \sigma^2 \left(\boldsymbol{V}^T\boldsymbol{V}\right)^{-1}\Big|_{\hat{\boldsymbol{\theta}}} \tag{4.28}$$

and

$$\Omega_{\hat{\boldsymbol{\theta}}} \approx \sigma^2 \left(\boldsymbol{V}^T\boldsymbol{V}\right)^{-1}\Big|_{\hat{\boldsymbol{\theta}}} + \frac{\partial^2 f}{\partial \boldsymbol{\theta}^2}(\boldsymbol{y} - f(\boldsymbol{x}, \boldsymbol{\theta}))\Big|_{\hat{\boldsymbol{\theta}}} \tag{4.29}$$

respectively. Although theoretically, the second term in (4.29) is zero, when the true values of the parameters are unknown and the sample is small, the second term may contribute to the estimate of $se(\hat{g})$ such that the values of the observed variance-covariance matrix do depend on which of (4.28) or (4.29) is used. We have found that when (4.28) is used as the basis for computing $se(\hat{g})$, and when the derivatives are computed numerically, the error in the estimate of $se(\hat{g})$ may be significant enough to cause the linear approximation reference line to *appear* to not be tangent to the profile t curve. Our experience is supported by that of Donaldson and Schnabel (1987) who investigated how linearization confidence intervals are affected by the manner in which the variance-covariance matrix is computed. They found that when the variance-covariance computations were based on analytic expressions for the required derivatives, little difference existed between the approximations; however, they noted

that using numerical methods to approximate the derivatives sometimes resulted in a serious degradation of the estimates.

4.6 Limitations of Profiling

Although profiling is a powerful method for finding likelihood intervals for functions of parameters, there are cases for which the method, whether based on optimization or reparameterization, is not appropriate. To appreciate the limitations of the algorithm, it is helpful to consider the method from both the optimization perspective and the reparameterization perspective.

A likelihood interval for a function $g(\boldsymbol{\theta})$ requires finding the maximum and minimum values of g over the likelihood region for $\boldsymbol{\theta}$, where the level of the critical contour is based on $F(1, n-p; \alpha)$. When g is linear and the likelihood region for $\boldsymbol{\theta}$ is a convex function, the maximum and minimum of g will lie on the boundary of the region. In this case the maximum and minimum values of g occur at the points where the the contours of g are tangent to the ellipse defining the confidence region for $\boldsymbol{\theta}$.

When g and/or the model are nonlinear, there are several possible cases:

1. g_{max} and g_{min} occur on the boundary of the likelihood region for $\boldsymbol{\theta}$, as for the linear case

2. g_{max} and/or g_{min} is/are located in the interior of the likelihood region for $\boldsymbol{\theta}$

3. g intersects the likelihood region for $\boldsymbol{\theta}$ at only a single point, and so the likelihood interval is a single point (i.e., there is only a single plausible value for g at the given level of confidence. Usually this implies that $g(\hat{\boldsymbol{\theta}})$ is undefined and that g is defined only at a single point which lies on the boundary of the likelihood region for $\boldsymbol{\theta}$.

4. g does not intersect with the likelihood region for $\boldsymbol{\theta}$. This case can occur if

the parameters of the fitted model are not appropriately constrained during the fitting of the model.

5. g has multiple local optima which fall on or within the likelihood region for $\boldsymbol{\theta}$.

These cases are illustrated in Figure 4.1 for a model involving two parameters. It is clear that nonlinearity significantly complicates the inference results. Often the likelihood intervals for a nonlinear $g(\boldsymbol{\theta})$ will be asymmetric about $g(\hat{\boldsymbol{\theta}})$, and for Case 5, the likelihood regions may be disjoint (Donaldson and Schnabel, 1987).

Consider now the mathematical implications of these various cases. For the reparameterization method, the equation set (4.9) requires the inverse of $g(\boldsymbol{\theta})$ to exist. That is, it must be possible to solve for θ_1 in terms of the vector of transformed variables $\boldsymbol{\phi}$. However, the solution for θ_1 need not necessarily be explicit . Typically, a numerical search scheme will easily produce a solution for θ_1. When θ_1 is left as an implicit function of $\boldsymbol{\phi}$, there may be multiple solutions. The non-uniqueness of the solution is not an issue if the likelihood interval for $\boldsymbol{\theta}$ includes only one of the solutions.

Computing a likelihood interval for $g(\boldsymbol{\theta})$ by the reparameterization method generally requires computing the derivatives of the likelihood function with respect to the new parameters. By (4.14), the inverse of $\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}$ must exist over the appropriate likelihood region for $\boldsymbol{\theta}$. This means that no two parameters can be defined such that they affect the function $f(\boldsymbol{x}, \boldsymbol{\theta})$ in exactly the same way.

The limits imposed on $g(\boldsymbol{\theta})$, as determined by the reparameterization approach, can also be motivated from the optimization perspective. As shown above, the optimization problem may be solved by the method of Lagrange multipliers. The necessary and sufficient conditions for $\tilde{\boldsymbol{\theta}}$ to be an optimum can be found in most textbooks on optimization (see, for example, Edgar and Himmelblau, 1988). The first necessary condition is that $f(\boldsymbol{x}, \boldsymbol{\theta})$ and $g(\boldsymbol{\theta})$ be twice differentiable at $\tilde{\boldsymbol{\theta}}$ for all c, where $\tilde{\boldsymbol{\theta}}$ is a solution of $g(\boldsymbol{\theta}) = c$. Thus, $f(\boldsymbol{x}, \boldsymbol{\theta})$ and $g(\boldsymbol{\theta})$, as well as their derivatives, must

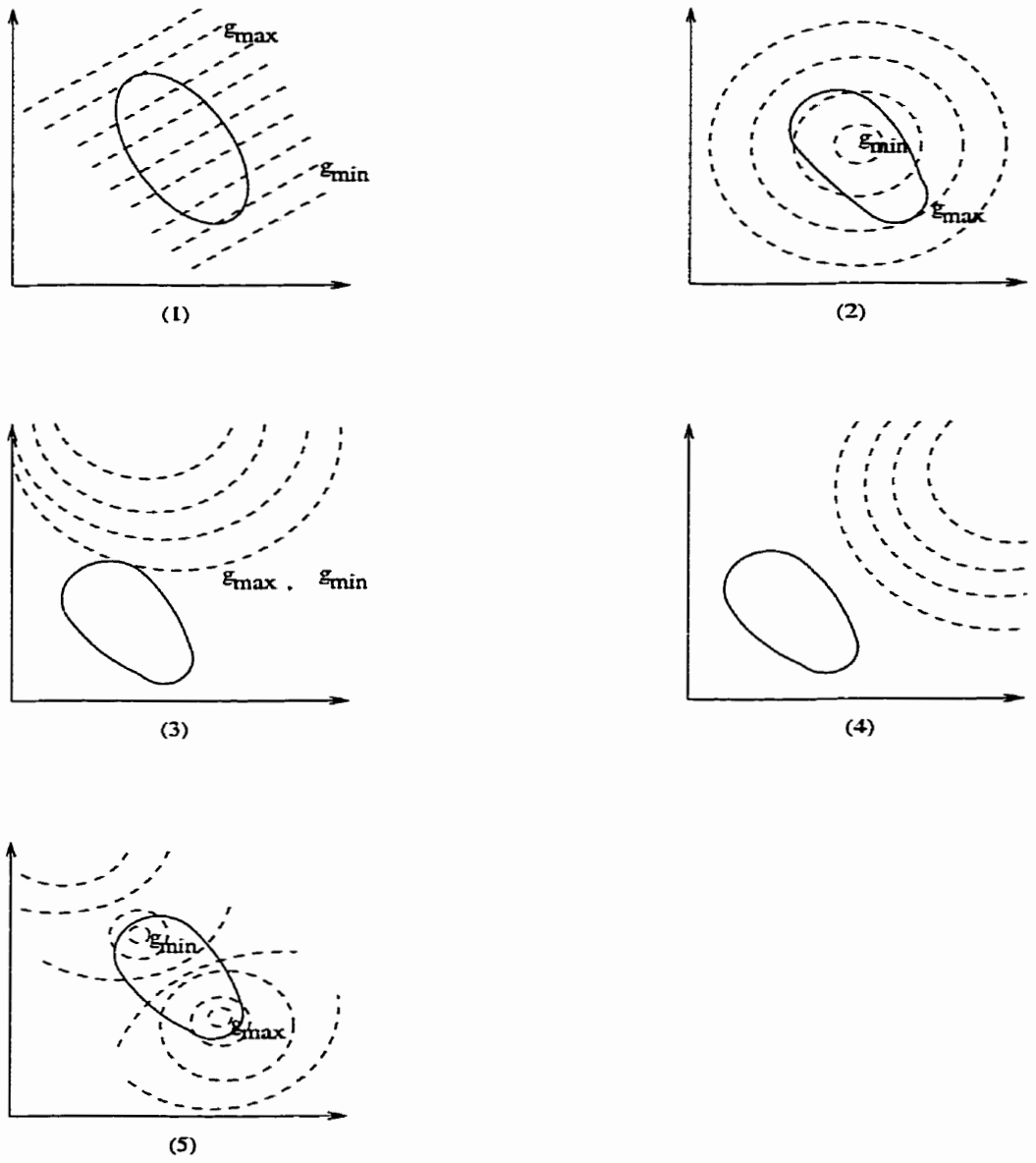

Figure 4.1: An illustration of various inference scenarios in two dimensions. The figures are plotted in the space of the parameters. KEY: - - contours of $g(\boldsymbol{\theta})$; – contours of $L(\boldsymbol{\theta})$.

exist at $\hat{\boldsymbol{\theta}}$. Therefore, profiling will fail for Cases 3 and 4 above because the profiling algorithm begins at $\hat{\boldsymbol{\theta}}$ where $g(\boldsymbol{\theta})$ and/or $f(\boldsymbol{\theta})$ do not exist.

The other necessary and sufficient conditions are either implied by the first, or deal with issues of inequality constraints, of which there are none in the profiling optimization problem. However, there is one further sufficient condition worthy of mentioning, namely that the Hessian matrix of the augmented Lagrangian equation be positive definite. Profiling may fail for Case 5 since the Hessian may not be positive definite. Also, note that for Case 5, profiling may fail to take account of multiple optima, basing a continuous likelihood region on one local optimum only.

We have discussed theoretical cases for which profiling is inappropriate. In our experience, these situations arise infrequently and are usually the result of data which contain insufficient information to estimate the parameters of the model, a model which does not adequately represent the system from which the data were sampled, or a poorly chosen parameterization of the model (for example, when the parameterization does not implicitly constrain the parameters to remain within a physically meaningful domain). In such cases, fitting the model to the data is inadvisable and a failure of the profiling algorithm is a signal that additional data and/or model reformulation are required.

The conditions described in this section, both from optimization and reparameterization perspectives, determine whether a solution exists. However, there is a special class of $g(\boldsymbol{\theta})$ for which a solution exists, but for which the profiling algorithm is not appropriate. Although profiling may be inappropriate for Cases 3 and 4, one might argue that the occurrence of these situations is not interesting (although important to document). However, there are practical and interesting cases related to the situation in Figure 4.1, Panel 2 which present problems for the profiling algorithm.

4.6.1 When the Likelihood Interval Exists but Profiling Fails

There exists a special class of functions $g(\boldsymbol{\theta})$ for which profiling is not an appropriate method for computing likelihood intervals. When $g(\boldsymbol{\theta})$ is not a monotonic function of $\boldsymbol{\theta}$, and when an unconstrained optimum of $g(\boldsymbol{\theta})$ is located at the same point as the unconstrained maximum of $L(\boldsymbol{\theta})$, the profiling algorithm will fail. In this case, the least squares estimate of $\boldsymbol{\theta}$ is also the location of the unconstrained optimum of $g(\boldsymbol{\theta})$, and consequently

$$\left.\frac{\partial g}{\partial \boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}} = 0 \tag{4.30}$$

and

$$\begin{aligned} se(\hat{g}) &= s\sqrt{\left.\frac{\partial g}{\partial \boldsymbol{\theta}}^T \left(\frac{\partial f}{\partial \boldsymbol{\theta}}^T \frac{\partial f}{\partial \boldsymbol{\theta}}\right)^{-1} \frac{\partial g}{\partial \boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}}} \\ &= 0 \end{aligned} \tag{4.31}$$

where $se(\hat{g})$ is the standard error of $g(\hat{\boldsymbol{\theta}})$. The profiling algorithm fails since its step sizes are based on multiples of $se(\hat{g})$. It is also not possible to determine the reference line since this is based on

$$\tau_{lin} = \frac{g(\tilde{\boldsymbol{\theta}}) - g(\hat{\boldsymbol{\theta}})}{se(\hat{g})} \tag{4.32}$$

When $se(\hat{g})$ is zero, this statistic can not be computed. Furthermore, the optimization algorithm proposed by Chen (1996) requires the calculation of the inverse of the

matrix:

$$\begin{pmatrix} \ddot{\boldsymbol{L}}(\hat{\boldsymbol{\theta}}) + \lambda_c \ddot{g}(\hat{\boldsymbol{\theta}}) & \dot{g}(\hat{\boldsymbol{\theta}}) \\ \dot{g}^T(\hat{\boldsymbol{\theta}}) & \mathbf{0} \end{pmatrix} \tag{4.33}$$

where a dot overstrike represents the derivative with respect to $\boldsymbol{\theta}$ and a double dot overstrike represents the second derivative. When $\dot{g}(\hat{\boldsymbol{\theta}}) = \frac{\partial g}{\partial \boldsymbol{\theta}}\Big|_{\hat{\boldsymbol{\theta}}} = 0$, this matrix is singular and the algorithm fails (Quinn et al., 1999a; Chapter 3).

To illustrate these ideas we construct a simple example. Suppose we had a two-parameter model for which $\hat{\boldsymbol{\theta}}^T = (1, 2)$. If $g(\boldsymbol{\theta}) = \theta_1(\theta_1 - 2)$, then it is straightforward to verify that the standard error of $g(\boldsymbol{\theta})$ is zero and profiling will fail. Although this is a constructed example, there are practical cases in which $g(\boldsymbol{\theta})$ and $f(\boldsymbol{x}, \boldsymbol{\theta})$ reach an optimum at the same location. In control engineering, a measure of controller performance is

$$g(\boldsymbol{\theta}) = performance = ||\boldsymbol{y}_{target} - \boldsymbol{y}_{actual}||_2 \tag{4.34}$$

Note that $\boldsymbol{y}_{actual}$ is fixed for a given realization of the system and does not depend on the estimated parameters of the system. In many applications, $\boldsymbol{y}_{target}$ is based on a model of the system such that

$$\boldsymbol{y}_{target} = h(\boldsymbol{y}(\boldsymbol{\theta})) \tag{4.35}$$

If $h(\cdot)$ is a monotonic function and if the same set of observations $\boldsymbol{y}_{actual}$ is used to estimate $\boldsymbol{\theta}$ as is used to compute performance, then profiling will fail because the location of the unconstrained minimum of $g(\boldsymbol{\theta})$ will be at $\hat{\boldsymbol{\theta}}$.

As another example, consider the Coefficient of Determination, R^2, where

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.36}$$

where $\tilde{y}_i$ represents the value of y computed on the basis of a model with parameter values $\tilde{\boldsymbol{\theta}}$. For regression models with additive error which is iid normal, the maximum likelihood estimates of the parameters are those which minimize the sum of squared residuals, i.e., those which minimize:

$$\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \tag{4.37}$$

From (4.36), R^2 depends only on the sum of squared residuals and that sample variance of y. Since the sample variance does not depend on the estimated parameters and is a fixed quantity for a given set of data, the values of the parameters which minimize the sum of squared residuals are also those which maximize R^2. In this way, $\hat{\boldsymbol{\theta}}$ is the location of an unconstrained optimum of R^2 and the profiling algorithm would fail for $g(\boldsymbol{\theta}) = R^2$.

The problems created for the profiling algorithm do *not* imply that a likelihood interval for $g(\theta)$ cannot be found. They simply suggest that an alternate method is required. To solve the problem, consider solving the alternate optimization problem:

$$\begin{aligned} &\text{Maximize} \quad g(\boldsymbol{\theta}) \\ &\text{Subject to} \quad L(\boldsymbol{\theta}) \geq L_{crit} \end{aligned} \tag{4.38}$$

Using any constrained optimization algorithm the upper limit of the likelihood interval for $g(\boldsymbol{\theta})$ is the constrained maximum found in (4.38). The lower limit of the likelihood interval for $g(\boldsymbol{\theta})$ is the minimum of $g(\boldsymbol{\theta})$ on or within the joint likelihood region for $\boldsymbol{\theta}$. For this special case, the unconstrained minimum of $g(\boldsymbol{\theta})$ lies within the likelihood

region for $\boldsymbol{\theta}$; therefore this must be the value of the lower limit of the likelihood interval for $g(\boldsymbol{\theta})$. It is independent of the value of the confidence level and requires no optimization. Note that while a likelihood interval can be found for these cases, joint likelihood regions can not.

Returning to the example of the Coefficient of Determination, a likelihood interval can be constructed using the minimization/maximization approach. Because the unconstrained optimum of R^2 occurs at $\hat{\boldsymbol{\theta}}$, it lies within all likelihood regions for the parameters. Therefore, $R^2|_{\hat{\boldsymbol{\theta}}}$ is the upper limit of the likelihood interval for R^2. To find the lower limit of R^2, solve the constrained optimization problem:

$$\begin{aligned} &\text{Minimize} && R^2 \\ &\text{Subject to} && ln(L(\boldsymbol{\theta})) \geq ln(L(\hat{\boldsymbol{\theta}})) - \frac{1}{2}F(1, n-p; \alpha) \end{aligned} \tag{4.39}$$

The constrained minimum of R^2 is the lower limit of the likelihood interval. For general nonlinear models, there is no analytic solution to the constrained optimization problem, and typically, numerical methods are employed.

4.7 Conclusion

When making decisions on the basis of a function of parameters $g(\boldsymbol{\theta})$ it is important to have a reliable measure of the uncertainty for any point estimate of $g(\boldsymbol{\theta})$. Generalized profiling provides a more reliable means of estimating likelihood intervals for functions of parameters than the commonly used linearization approach.

There are two approaches to generalizing the profiling algorithm. One is based on constrained optimization (Chen, 1991; Chen and Jennrich, 1996), and the other on reparameterization (Clarke, 1987; Bates and Watts, 1988; Ross, 1990). The equivalence of the two approaches has been shown.

By considering generalized profiling from both perspectives, the merits and limita-

tions of the profiling algorithm have been discussed, and cases for which the algorithm fails have been identified. An alternative approach based on minimizing and maximizing $g(\boldsymbol{\theta})$ over a likelihood region, has been proposed for one of these cases.

4.8 Acknowledgements

The authors gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada and the School of Graduate Studies of Queen's University.

4.9 Nomenclature

c	= a constant
$D^2(\boldsymbol{\theta})$	= the deviance
$\boldsymbol{e}$	= $n \times 1$ column vector of estimated random errors
$F(p, n-p; \alpha)$	= upper α quantile for the F distribution with p and $n-p$ degrees of freedom
$g(\boldsymbol{\theta})$	= a function of parameters
$g^{-1}(\boldsymbol{\theta})$	= the first element of the inverse of the reparameterization involving $g(\boldsymbol{\theta})$
$\hat{\dot{g}}$	= vector of derivative of $g(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ evaluated at $\hat{\boldsymbol{\theta}}$
$L(\boldsymbol{\theta})$	= likelihood function evaluated at $\boldsymbol{\theta}$
$\ddot{L}(\hat{\boldsymbol{\theta}})$	= matrix of second derivatives of the likelihood function with respect to $\boldsymbol{\theta}$ evaluated at $\hat{\boldsymbol{\theta}}$
$LR(c)$	= likelihood ratio for testing the null hypothesis that $g(\boldsymbol{\theta}) = c$

n = number of observations

p = number of estimated parameters

s = estimated standard deviation of the random errors

$S(\boldsymbol{\theta})$ = sum of squared errors

se = standard error

$t(n-p;\alpha/2)$ = upper $\alpha/2$ quantile of the t distribution with $n-p$ degrees of freedom

$\boldsymbol{x}$ = $1 \times m$ row vector of m independent variables

$\boldsymbol{X}$ = $n \times p$ matrix of elements x_{ij} representing the level of the j^{th} independent variable for observation i

y = response variable

$\boldsymbol{y}$ = $n \times 1$ column vector of values of the response variable

$\boldsymbol{V}$ = $n \times p$ matrix of elements v_{ij} representing the first derivative of $f(\boldsymbol{x}_i, \boldsymbol{\theta})$ with respect to the j^{th} parameter

$z^*(c)$ = signed root deviance (SRD)

Greek letters

α = significance level

ϵ = additive random error

$\boldsymbol{\epsilon}$ = $n \times 1$ column vector of random errors

θ_q = q^{th} parameter of a model

$\boldsymbol{\theta}$ = $p \times 1$ column vector of parameters of a model

λ_i = i^{th} Lagrange multiplier

σ = standard deviation

$\tau(g(\boldsymbol{\theta}))$ = profile t statistic for $g(\boldsymbol{\theta})$

ϕ_q = q^{th} parameter of a reparameterized model

$\boldsymbol{\phi}$	$= p \times 1$ column vector of an alternate set of parameters of a model
$\chi^2(1)$	= the chi-squared distribution with 1 degree of freedom
$\boldsymbol{\Omega_\theta}$	= variance-covariance matrix for $\hat{\boldsymbol{\theta}}$

Superscripts

$*$	= a true value
$\hat{}$	= a maximum likelihood estimate
$\tilde{}$	= a constrained estimate

Abbreviations

iid	independently and identically distributed
MLE	maximum likelihood estimate

Chapter 5

Use of Expected Profiling for Likelihood Interval Prediction in Time Series Models

5.1 Abstract

When collecting data for the purpose of fitting a time series model, it is of interest to know how much data will be "enough" to satisfy the estimation objectives. We develop a tool called an n-plot which indicates how the likelihood limits for individual parameters (or functions of parameters) of ARMA models are expected to change as the length of the times series data set increases. The n-plots are based on expected profiles which, in themselves, provide useful information prior to data collection. The basis for the expected profiles is an expression for the expected value of the likelihood ratio function. The use and value of expected profiling and n-plots are illustrated with two examples. The n-plot can be used to design the length of a dynamic experiment and to estimate how long a time series needs to be in order to obtain "reliable" estimates of the parameters or functions of parameters of a time series model.

5.2 Introduction

The asymptotic statistical properties of estimates of parameters and forecasts from fitted time series models have received considerable attention (Rao, 1962; Yamamoto, 1976; Ljung, 1985; Taniguchi, 1986). Asymptotic expressions are applicable when the number of observations n in a realization of a time series is "large"; however, the issue of how large n must be to satisfy estimation objectives is rarely addressed (Eliason, 1993). We show that a value of n which is "large enough" depends on the model form, its parameterization, and the values of the parameters themselves, where "large enough" means that the Cramer-Rao lower bounds are expected to provide a close approximation to the likelihood limits.

The issue of how much data should be collected in order to obtain a useful fitted process model is an important topic in dynamic design of experiments (Isermann, 1980). Åström (1980) noted that although it is well known that maximum likelihood estimates have good asymptotic properties, the extent to which these properties are achieved when small data sets are used is unclear. This paper addresses these issues by developing expressions for *expected profiles*. An expected profile is analogous to the profile t plot introduced by Bates and Watts (1988), but do not depend on the availability of a set of data. Expected profiles provide useful information to experimenters prior to data collection about how much data will be needed to provide useful estimates of parameters and functions of parameters in time series models.

Profiling (Bates and Watts, 1988; Chen, 1991; Chen and Jennrich, 1996) is a likelihood ratio approach to estimating uncertainties in parameters and functions of parameters in proposed models. It has been used to estimate likelihood intervals for parameters in steady-state models (Bates and Watts, 1988; Chen, 1991; Watts, 1994) and in ARMA models (Lam and Watts, 1991; Chen and Jennrich, 1996). Because profiling is a likelihood-based method, its results are specific to the particular model and data set under consideration. However, in the case of ARMA models, the model

itself defines how the process is expected to evolve over time and it is possible to calculate expected values for some of the properties of the model *a priori* to any data collection (i.e., in the absence of a realization of the process). This is also possible with other classes of models, but we will focus on time series models.

We derive an expression for the expected value of the likelihood ratio for the hypothesis $g(\boldsymbol{\theta}) = c$, relative to any specified alternative hypothesis, where $\boldsymbol{\theta}$ is the vector of parameters in a proposed time series model, $g(\boldsymbol{\theta})$ is a function of those parameters, and c is a constant. Based on this expression, expected profiles for individual parameters or a function of parameters $g(\boldsymbol{\theta})$ are constructed and used to judge how long a time series needs to be in order to obtain "reliable" estimates of the parameters $\boldsymbol{\theta}$ or a function of parameters $g(\boldsymbol{\theta})$.

Profile t plots for individual parameters, or functions of the parameters, in a proposed model depend on the form of the model, its parameterization, and the data themselves. Anomalies in the data may manifest themselves as an increase or decrease in observed nonlinear behavior of parameters or functions of parameters. Expected profiles are computed in the absence of data, and therefore do not capture any of the uncertainty and nonlinearity associated with a particular data set. However, the expected profiles do provide important information about the inherent uncertainty and nonlinearity associated with the model and its parameterization. Note that the uncertainty and nonlinearity associated with a model are functions of the true values of the parameters; therefore expected profiles require assumptions about the values of the parameters, as well as about the form of the model. The usefulness of an expected profile is in the information it provides about the behaviour of the level of uncertainty associated with a parameter or function of parameters to be expected as n, the length of the time series, changes.

The paper proceeds as follows. We begin by reviewing the theory of profiling. Then, an expression for the expected value of the profiling statistic τ^2 is developed

for the case of ARMA models. To illustrate the use of this result, profile t plots for the parameters in two fitted models are shown and compared to the expected profiles for those parameters. We examine how the expected likelihood intervals for these two examples change as a function of n. Next, an example of computing expected profiles for a function of parameters is presented. We conclude with some ideas for future work.

5.3 Profiling

Profiling (Bates and Watts, 1988; Chen, 1991; Lam and Watts, 1991; Severini and Staniswalis, 1994; Chen and Jennrich, 1996, Quinn et al., 1999a; Chapter 3) is a graphical means for displaying inference results for parameters, and functions of parameters, of proposed models. Bates and Watts (1988) developed the algorithm specifically to summarize inferential results for parameters of nonlinear regression models. The method is based on the fact that, for regression models which are linear in the parameters and for which the additive errors are independently and identically normally distributed (iid $N(0, \sigma^2)$),

$$\frac{\theta_i - \hat{\theta}_i}{se(\hat{\theta}_i)} = sign(\theta_i - \hat{\theta}_i)\sqrt{\frac{S(\tilde{\boldsymbol{\theta}}) - S(\hat{\boldsymbol{\theta}})}{s^2}} \tag{5.1}$$

and the left and right hand sides have a t distribution with $(n-p)$ degrees of freedom, where θ_i is the parameter being investigated, $\hat{\theta}_i$ is the maximum likelihood (least squares) estimate of θ_i, $se(\hat{\theta}_i)$ is the standard error of $\hat{\theta}_i$, $S(\hat{\boldsymbol{\theta}})$ is the minimum sum of squares of residuals, $S(\tilde{\boldsymbol{\theta}})$ is the sum of squares of residuals when the vector of unknown parameters $\boldsymbol{\theta}$ is equal to $\tilde{\boldsymbol{\theta}}$, and $\tilde{\boldsymbol{\theta}}$ is the vector of parameter values which minimizes the sum of squares of residuals given the constraint that $\theta_i = c$, where c is

a constant.

$$s^2 = \frac{S(\hat{\boldsymbol{\theta}})}{n-k} \tag{5.2}$$

is an estimate of σ^2 with $(n-k)$ degrees of freedom. A $(1-\alpha)100$ % confidence interval for θ_i is

$$-t(n-k;\alpha/2) \leq \delta(\theta_i) \leq t(n-k;\alpha/2) \tag{5.3}$$

where

$$\delta(\theta_i) = \frac{\theta_i - \hat{\theta}_i}{se(\hat{\theta}_i)} \tag{5.4}$$

and $t(n-k;\alpha/2)$ denotes the upper $\alpha/2$ quantile of the t distribution with $n-k$ degrees of freedom, n is the number of observations to which the proposed model has been fitted, and k is the number of estimated parameters in the proposed model. For nonlinear models, it is common to see nominal $(1-\alpha)100$ % confidence intervals for the parameters approximated by (5.3). Although this linearization approach is computationally simple, it has been shown to be unreliable for some models (including time series models) that are nonlinear in the parameters (Ansley and Newbold, 1979; Donaldson and Schnabel, 1987; Bates and Watts, 1988; Lam and Watts, 1991). The idea underlying profiling is that inferences about the parameters of nonlinear models would be more accurate if they were based on

$$\tau(\theta_i) = sign(\theta_i - \hat{\theta}_i)\sqrt{\frac{S(\tilde{\boldsymbol{\theta}}) - S(\hat{\boldsymbol{\theta}})}{s^2}} \tag{5.5}$$

because this expression takes account of the nonlinearity of the sums of squares surface. Indeed, Donaldson and Schnabel (1987) and Chen (1991) have shown that, for

the case of nonlinear regression models, the coverage probability of likelihood intervals based on (5.5), i.e., based on profiling, is consistently closer to the nominal confidence level than that of linearization intervals.

Lam and Watts (1991) extended the theory of profiling to encompass time series models using a modified sum of squares appropriate for such models. Chen (1991) and Chen and Jennrich (1996) developed a theory of profiling in terms of likelihood ratios and constrained optimization. This formulation of the profiling algorithm is very general and may be used for several classes of models, including time series models. Furthermore, the constrained optimization approach facilitates computing inference results for functions of parameters.

In Section 5.4 the theory of expected profiling is developed in terms of the likelihood function for ARMA models. In this section we focus on the profiling algorithm as it was developed by Chen (1991) and Chen and Jennrich (1996).

Note that $\tau(\theta_i)$ may be written, equivalently but more generally, in terms of a function of the likelihood ratio:

$$\tau(\theta_i) = sign(\theta_i - \hat{\theta}_i)\sqrt{-2\ln\frac{L(\tilde{\boldsymbol{\theta}})}{L(\hat{\boldsymbol{\theta}})}} \tag{5.6}$$

where $L(\boldsymbol{\theta})$ is the likelihood function for the parameters of a specified model. This expression for $\tau(\theta_i)$ is very general in that it may be used to obtain inferences about individual parameters of any model so long as an expression for the likelihood function can be found. Note that to compute $\tau(\theta_i)$ the solution $\tilde{\boldsymbol{\theta}}$ to the following constrained optimization problem must first be obtained:

Maximize:

$$L(\boldsymbol{\theta}) \tag{5.7}$$

Subject to:

$$\theta_i = c$$

To construct a profile t plot for an individual parameter θ_i, the optimization problem in (5.7) is solved for a series of values of c greater than and less than the maximum likelihood estimate of θ_i, and the corresponding values of $\tau(\theta_i)$ are plotted against the values of θ_i. A marginal $(1-\alpha)\%$ likelihood interval for θ_i then includes all values of θ_i such that:

$$-t(n-k;\alpha/2) \leq \tau(\theta_i) \leq t(n-k;\alpha/2) \tag{5.8}$$

The limits of the interval can be obtained directly from the profile t plot by finding the values of θ_i at which horizontal lines extending from $\tau = \pm t(n-k;\alpha/2)$ intersect with the profile curve. This follows from the fact that, under the null hypothesis, $\theta_i = c$, the log likelihood ratio:

$$2\ln(LR) = 2[\ln L(\hat{\boldsymbol{\theta}}) - \ln L(\tilde{\boldsymbol{\theta}})] \tag{5.9}$$

has an asymptotic χ^2 distribution with one degree of freedom (Ravishanker et al., 1990). Thus, $\sqrt{2\ln(LR)}$ is asymptotically normally distributed. We use the student t distribution in place of the normal distribution to account for the estimation of the noise variance.

In many cases, a proposed model will be used to compute the value of a function of parameters $g(\boldsymbol{\theta})$. For example, a proposed model might be used to make predictions which are simply functions of the parameters of the proposed model. Likelihood intervals for a function of parameters $g(\boldsymbol{\theta})$ can be obtained by following the procedure outlined above, but in this case the constrained optimization problem becomes:

Maximize:

$$L(\boldsymbol{\theta}) \tag{5.10}$$

subject to:

$$g(\boldsymbol{\theta}) = c$$

and the expression for τ is:

$$\tau(g(\boldsymbol{\theta})) = sign(g(\boldsymbol{\theta}) - g(\hat{\boldsymbol{\theta}}))\sqrt{-2\ln\frac{L(\tilde{\boldsymbol{\theta}})}{L(\hat{\boldsymbol{\theta}})}} \tag{5.11}$$

The profile t plots are then plots of $\tau(g(\boldsymbol{\theta}))$ versus $g(\boldsymbol{\theta})$. Profiling an individual parameter θ_i is simply the special case where $g(\boldsymbol{\theta}) = \theta_i$.

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 inference results would be. A reference line, $\delta(g(\boldsymbol{\theta}))$ versus $g(\boldsymbol{\theta})$, is typically included on profile t plots. This reference line may be used to obtain the linearization confidence intervals for $g(\boldsymbol{\theta})$ and to judge the relative curvature of the function of parameters (Chen, 1991).

5.4 Expected Profiling

Consider an ARMA(p,q) model of the form:

$$\phi(B)y_t = \theta(B)a_t \tag{5.12}$$

where $\phi(B) = (1 + \phi_1 B + \ldots + \phi_p B^p)$, $\theta(B) = (1 + \theta_1 B + \ldots + \theta_q B^q)$, B is the backshift operator defined as $By_t = y_{t-1}$, $\{y_t\}$ is an observed stationary stochastic process with mean zero, and $\{a_t\}$ is a normally distributed white noise process such

that all elements of a_t are independently and identically normally distributed (i.e., $\{a_t\}$ is a sequence of iid $N(0, \sigma_a^2)$ random variables). For a time series model, $\boldsymbol{\theta}^T = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$. Although σ_a^2 is usually unknown, we can estimate it from the residuals of the fitted model and we do not include it in $\boldsymbol{\theta}$. Note that a nonstationary time series or one having a non-zero mean may be transformed to conform to the above model by first appropriately differencing or mean centering the data, respectively.

Building on the work of Chen (1991), Chen and Jennrich (1996), and Lam and Watts (1991) for profiling times series models, we develop a theory for constructing expected profiles, which are profile t plots for functions of the parameters of ARMA models constructed in the absence of data.

To create expected profile t plots we first develop an expression for:

$$\begin{aligned} E\{\tau^2\} &= E\left\{-2ln\left(\frac{L(\tilde{\boldsymbol{\theta}})}{L(\hat{\boldsymbol{\theta}})}\right)\right\} \\ &= 2\left[E\{\ln L(\hat{\boldsymbol{\theta}})\} - E\{\ln L(\tilde{\boldsymbol{\theta}})\}\right] \\ &= 2\left[E\{\mathcal{L}(\hat{\boldsymbol{\theta}})\} - E\{\mathcal{L}(\tilde{\boldsymbol{\theta}})\}\right] \end{aligned} \tag{5.13}$$

where $\mathcal{L}(\boldsymbol{\theta})$ is the natural logarithm of the likelihood function of $\boldsymbol{\theta}$. Therefore, we require expressions for the expected values of $\mathcal{L}(\hat{\boldsymbol{\theta}})$ and $\mathcal{L}(\tilde{\boldsymbol{\theta}})$.

Since we assume that the a_t's are iid $N(0, \sigma_a^2)$, the likelihood function for $\boldsymbol{\theta}$ given the data $\boldsymbol{y}_n = \{y_1, y_2, \ldots, y_n\}$ is

$$L(\boldsymbol{\theta}|\boldsymbol{y_n}) = (2\pi\sigma_a^2)^{-n/2}|\boldsymbol{\Omega}_n|^{-1/2}\exp\left(\frac{-\boldsymbol{y_n}'\boldsymbol{\Omega}_n^{-1}\boldsymbol{y_n}}{2\sigma_a^2}\right) \tag{5.14}$$

where $L(\boldsymbol{\theta}|\boldsymbol{y_n})$ is the likelihood function for the unknown parameters of the model, and $\sigma_a^2\boldsymbol{\Omega}_n$ denotes the $n \times n$ variance-covariance matrix for $\boldsymbol{y_n}$ as specified by the model (Box and Jenkins, 1976). The use of the subscript n follows the notation of Box and Jenkins (1976) and is used to emphasize the dependence of the likelihood

function on the length of the time series. Hereafter, the abbreviated notation $L(\boldsymbol{\theta})$ will be used since the focus of this work is on the dependency of the likelihood on the unknown parameters. Substituting $S(\boldsymbol{\theta}) = \boldsymbol{y_n}'\boldsymbol{\Omega}_n^{-1}\boldsymbol{y_n}$ into (5.14), and taking natural logarithms,

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{n}{2}ln(2\pi\sigma_a^2) - \frac{1}{2}\ln|\boldsymbol{\Omega}_n| - \frac{S(\boldsymbol{\theta})}{2\sigma_a^2} \tag{5.15}$$

Expressions for $\mathcal{L}(\hat{\boldsymbol{\theta}})$ and $\mathcal{L}(\tilde{\boldsymbol{\theta}})$ are then as follows:

$$\mathcal{L}(\hat{\boldsymbol{\theta}}) = -\frac{n}{2}ln(2\pi\sigma_a^2) - \frac{1}{2}\ln|\hat{\boldsymbol{\Omega}}_n| - \frac{S(\hat{\boldsymbol{\theta}})}{2\sigma_a^2} \tag{5.16}$$

and

$$\mathcal{L}(\tilde{\boldsymbol{\theta}}) = -\frac{n}{2}ln(2\pi\sigma_a^2) - \frac{1}{2}\ln|\tilde{\boldsymbol{\Omega}}_n| - \frac{S(\tilde{\boldsymbol{\theta}})}{2\sigma_a^2} \tag{5.17}$$

Both $S(\boldsymbol{\theta})$ and $\boldsymbol{\Omega}_n$ are functions of $\boldsymbol{\theta}$ and we explicitly distinguish them as being calculated based on $\tilde{\boldsymbol{\theta}}$ or $\hat{\boldsymbol{\theta}}$. The natural logarithm of the likelihood ratio is:

$$\begin{aligned} \mathcal{LR} &= \mathcal{L}(\tilde{\boldsymbol{\theta}}) - \mathcal{L}(\hat{\boldsymbol{\theta}}) \\ &= \frac{1}{2}\ln\left(\frac{|\hat{\boldsymbol{\Omega}}_n|}{|\tilde{\boldsymbol{\Omega}}_n|}\right) + \frac{1}{2\sigma_a^2}\left(S(\hat{\boldsymbol{\theta}}) - S(\tilde{\boldsymbol{\theta}})\right) \end{aligned} \tag{5.18}$$

Therefore,

$$\tau^2 = \frac{1}{\sigma_a^2}\left(S(\tilde{\boldsymbol{\theta}}) - S(\hat{\boldsymbol{\theta}})\right) + \ln\left(\frac{|\tilde{\boldsymbol{\Omega}}_n|}{|\hat{\boldsymbol{\Omega}}_n|}\right)$$

We are interested in computing $E\{\tau^2\}$. This amounts to finding expressions for $E\{S(\tilde{\boldsymbol{\theta}})\}$ and $E\{S(\hat{\boldsymbol{\theta}})\}$. Note that the values of the elements of the covariance matrices are fully defined by the model and the values of the parameters. If $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$, where $\boldsymbol{\theta}^*$ is the true value of $\boldsymbol{\theta}$, and if the starting values for the time series are known, then

it can be shown (Box and Jenkins, 1976) that:

$$S(\hat{\boldsymbol{\theta}}) = \sum_{t=1}^{n} a_t^2 \tag{5.19}$$

Therefore, under this assumption

$$\begin{aligned} E\{S(\hat{\boldsymbol{\theta}})\} &= \sum_{t=1}^{n} E\{a_t^2\} \\ &= n\sigma_a^2 \end{aligned} \tag{5.20}$$

However, when $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$,

$$\boldsymbol{y_n}'\boldsymbol{\Omega}_n^{-1}\boldsymbol{y_n} \neq \sum_{t=1}^{n} a_t^2 \tag{5.21}$$

It is therefore necessary to develop an expression for $S(\tilde{\boldsymbol{\theta}})$ which will take into account the increase in variance of the error terms when $\tilde{\boldsymbol{\theta}} \neq \boldsymbol{\theta}^*$.

Assume the true process can be represented by the ARMA model:

$$y_t = \frac{\theta^*(B)}{\phi^*(B)} a_t \tag{5.22}$$

where the superscript $*$ represents true values of the parameters. If a fitted model for the process based on a realization $\{y_t\}$ has parameter estimates $\hat{\boldsymbol{\theta}}$, the residuals from that fitted model are:

$$e_t = \frac{\hat{\phi}(B)}{\hat{\theta}(B)} y_t \tag{5.23}$$

When $\frac{\hat{\theta}(B)}{\hat{\phi}(B)} = \frac{\theta^*(B)}{\phi^*(B)}$, $e_t = a_t$ and the expected value of the log likelihood function is:

$$E\{\mathcal{L}(\hat{\boldsymbol{\theta}})\} = E\{\mathcal{L}(\boldsymbol{\theta}^*)\} = -\frac{n}{2} ln(2\pi\sigma_a^2) - \frac{1}{2}\ln|\boldsymbol{\Omega}_n^*| - \frac{n}{2} \tag{5.24}$$

In developing an expression for $E\{\mathcal{L}(\tilde{\boldsymbol{\theta}})\}$, we presume that the process can be modeled as

$$y_t = \frac{\tilde{\theta}(B)}{\tilde{\phi}(B)}\tilde{a}_t \tag{5.25}$$

But because y_t is actually a realization of (5.22),

$$\tilde{a}_t = \frac{\theta^*(B)}{\phi^*(B)}\frac{\tilde{\phi}(B)}{\tilde{\theta}(B)}a_t \tag{5.26}$$

Then,

$$\begin{aligned} E\{S(\tilde{\boldsymbol{\theta}})\} &= E\{\boldsymbol{y_n}'\tilde{\boldsymbol{\Omega}}_n{}^{-1}\boldsymbol{y_n}\} \\ &= tr\left(\tilde{\boldsymbol{\Omega}}_n{}^{-1}E\{\boldsymbol{y_n}\boldsymbol{y_n}'\}\right) \text{ (Mathai and Provost, 1992)} \\ &= \sigma_a^2 tr\left(\tilde{\boldsymbol{\Omega}}_n{}^{-1}\boldsymbol{\Omega}_n^*\right) \end{aligned} \tag{5.27}$$

Finally, we can write, for an unbiased model,

$$\begin{aligned} E\{\tau^2\} &= \frac{1}{\sigma_a^2}\left(\sigma_a^2 tr\left(\tilde{\boldsymbol{\Omega}}_n{}^{-1}\boldsymbol{\Omega}_n^*\right) - n\sigma_a^2\right) + \ln\left(\frac{\left|\tilde{\boldsymbol{\Omega}}_n\right|}{\left|\boldsymbol{\Omega}_n^*\right|}\right) \\ &= tr\left(\tilde{\boldsymbol{\Omega}}_n{}^{-1}\boldsymbol{\Omega}_n^*\right) - n + \ln\left(\frac{\left|\tilde{\boldsymbol{\Omega}}_n\right|}{\left|\boldsymbol{\Omega}_n^*\right|}\right) \end{aligned} \tag{5.28}$$

It is interesting to note that the expected value of τ^2 is independent of the variance of the white noise driving force σ_a^2. The vector $\tilde{\boldsymbol{\theta}}$ represents the conditional maximum likelihood estimates of the parameters given the constraint $g(\boldsymbol{\theta}) = c$. When a data set is available, $\tilde{\boldsymbol{\theta}}$ is found by solving the constrained optimization problem given in (5.10). When computing $E\{\tau^2\}$ we find $\tilde{\boldsymbol{\theta}}$ by minimizing $E\{\mathcal{L}(\boldsymbol{\theta})\}$ subject to the constraint $g(\boldsymbol{\theta}) = c$.

Now we have an expression for the expected value of τ^2 based on the full likelihood

ratio. To find the expected profile t plot of a function of parameters of an ARMA model for any value of n, we compute

$$\tau_{exp} = sign(g - \hat{g})\sqrt{E\{\tau^2\}} \tag{5.29}$$

The development of the expression for $E\{\tau^2\}$ is based on determining the values of $\tilde{\boldsymbol{\theta}}$. These values, in turn, are determined by the vectgor of true values $\boldsymbol{\theta}^*$, the form of the model and the constraint $g(\boldsymbol{\theta}) = c$. All of the standard assumptions about the random error driving force were made. Therefore, we assume the distribution of $E\{\tau^2\}$ is $F(1, n-p; \alpha)$, as is the case for τ^2 based on observed data. A $(1-\alpha)100\%$ expected likelihood interval for $g(\boldsymbol{\theta})$ includes all values of $g(\boldsymbol{\theta})$ such that $F(1, n-p; \alpha/2) \leq E\{\tau^2\} \leq F(1, n-p; 1-\alpha/2)$. Equivalently, the expected likelihood interval can be computed based on τ_{exp} since $\sqrt{F(1, n-p; \alpha)} = t(n-p; \alpha/2)$. An expected profile plot of a function of parameters $g(\boldsymbol{\theta})$ is a plot of τ_{exp} versus $g(\boldsymbol{\theta})$ or $\delta(g)$, where

$$\delta(g) = \frac{g(\boldsymbol{\theta}) - \hat{g}}{se(g(\hat{\boldsymbol{\theta}}))} \tag{5.30}$$

and

$$se(g(\hat{\boldsymbol{\theta}})) = \sqrt{\frac{\partial g}{\partial \boldsymbol{\theta}}^T \left[Cov(\hat{\boldsymbol{\theta}})\right]_{ii} \frac{\partial g}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}} \tag{5.31}$$

where $[Cov(\hat{\boldsymbol{\theta}})]_{ii}$ denotes the i^{th} diagonal element of the variance-covariance matrix for $\hat{\boldsymbol{\theta}}$ (Wei, 1990). This variance-covariance matrix is computed from the Cramer-Rao lower bound (Ljung, 1987):

$$Cov(\hat{\boldsymbol{\theta}}) \geq \sigma_a^2 [E\{\mathcal{I}\}]^{-1} \tag{5.32}$$

$$\geq \sigma_a^2 \left[\sum_{t=1}^{n} E\left\{\psi(t, \boldsymbol{\theta}^*)\psi^T(t, \boldsymbol{\theta}^*)\right\}\right]^{-1} \tag{5.33}$$

where $\mathcal{I}$ is the information matrix with elements:

$$\mathcal{I}_{ij} = \left. \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}, \sigma_a^2)}{\partial \theta_i \partial \theta_j} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \tag{5.34}$$

and

$$\psi(t, \boldsymbol{\theta}^*) = \left. \frac{\partial}{\partial \boldsymbol{\theta}} a(t, \boldsymbol{\theta}) \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \tag{5.35}$$

Linearization likelihood limits for $g(\boldsymbol{\theta})$ are given by:

$$g(\hat{\boldsymbol{\theta}}) \pm t(n-k; \alpha/2) se(g(\hat{\boldsymbol{\theta}})) \tag{5.36}$$

For a fixed value of n, it is convenient to plot profile t plots as τ versus $g(\theta)$. However, when plotting multiple profiles corresponding to changing values of n on one plot, it is more informative to plot $\tau(g)$ versus $\delta(g)$ in order that all profiles may be compared against the same reference line. When plotting τ versus $g(\boldsymbol{\theta})$, the slope of the reference line, $1/se(g(\hat{\boldsymbol{\theta}}))$, will change as n changes since $se(g(\hat{\boldsymbol{\theta}}))$ is a function of n. By plotting $\tau(g)$ versus $\delta(g)$, the dependency on n is removed since $\delta(g)$ is scaled by $se(g(\hat{\boldsymbol{\theta}}))$ and the reference line is always a straight line with unit slope passing through the origin (Bates and Watts, 1988).

By computing expected likelihood intervals for individual parameters θ_i (or for any function $g(\boldsymbol{\theta})$) over a range of values of n, we can then construct an n-plot, which is a plot of n versus the limits of the expected likelihood intervals. The intervals based on the Cramer-Rao lower bounds can also be shown on an n-plot for comparison. The n-plot provides a clear indication of how the uncertainty in an estimate of interest ($\hat{\theta}_i$ or $g(\hat{\boldsymbol{\theta}})$) can be expected to change as the length of the observed time series changes.

In order to compute values for τ_{exp}, indeed to develop the methodology of expected profiling itelf, several assumptions are required. One must choose a form for the model

and assume that it can represent the behaviour of the true system. Furthermore, one must choose values for the parameters and assume that they are equal to the "true" values of the parameters of the system. Although it seems unreasonable to think that al of these things can be known *a priori* to experimentation, good guesses can be made on the basis of work reported in the literature or engineering experience. The catch 22 of having to know the model and the values of its parameters in order to design an experiment to estimate the parameters is a feature of all nonlinear design of experiments work. Despite this, many nonlinear designs have proven useful and valuable. In fact, the Cramer-Rao bounds have been used for decades and these require the same guesses and assumptions as does expected profiling.

5.4.1 Computational Issues

For an ARMA(p,q) model, the expressions for the derivatives of a_t with respect to the parameters are (Ravishanker, 1994)

$$\frac{\partial a_t}{\partial \phi_k} = -\frac{1}{\phi(B)} a_{t-k} = -\frac{1}{\theta(B)} y_{t-k} \tag{5.37}$$

$$\frac{\partial a_t}{\partial \theta_l} = \frac{1}{\theta(B)} a_{t-l} = \frac{\phi(B)}{\theta^2(B)} y_{t-l} \tag{5.38}$$

Let

$$v(u) = cov\left(\frac{\partial a_t}{\partial \phi_k}, \frac{\partial a_{t+u}}{\partial \phi_k}\right) = cov\left(\frac{\partial a_t}{\partial \phi_k}, \frac{\partial a_t}{\partial \phi_{k+u}}\right) \tag{5.39}$$

$$\nu(u) = cov\left(\frac{\partial a_t}{\partial \theta_l}, \frac{\partial a_{t+u}}{\partial \theta_l}\right) = cov\left(\frac{\partial a_t}{\partial \theta_l}, \frac{\partial a_t}{\partial \theta_{l+u}}\right) \tag{5.40}$$

(Åström, 1980) and

$$\varrho_{\phi\theta}(u) = cov\left(\frac{\partial a_t}{\partial \phi_k}, \frac{\partial a_{t+u}}{\partial \theta_l}\right) \tag{5.41}$$

Note that

$$\varrho_{\phi\theta}(u) \neq \varrho_{\theta\phi}(u) \tag{5.42}$$

but

$$\varrho_{\phi\theta}(u) = \varrho_{\theta\phi}(-u) \tag{5.43}$$

Then, in order to calculate the lower bound for $Cov(\hat{\boldsymbol{\theta}})$ based on (5.33) so as to obtain a value for $se(g(\hat{\boldsymbol{\theta}}))$, we develop the expression:

$$\begin{aligned} \Upsilon &= E\{\mathcal{I}\} \qquad (5.44)\\ &= E\left\{\boldsymbol{\psi}\boldsymbol{\psi}^{\boldsymbol{T}}\right\} \\ &= \begin{bmatrix} \upsilon(0) & \cdots & \upsilon(p-1) & \varrho_{\phi\theta}(0) & \cdots & \varrho_{\phi\theta}(q-1) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \upsilon(p-1) & \cdots & \upsilon(0) & \varrho_{\phi\theta}(1-p) & \cdots & \varrho_{\phi\theta}(0) \\ \varrho_{\phi\theta}(0) & \cdots & \varrho_{\phi\theta}(1-p) & \nu(0) & \cdots & \nu(q-1) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \varrho_{\phi\theta}(q-1) & \cdots & \varrho_{\phi\theta}(0) & \nu(q-1) & \cdots & \nu(0) \end{bmatrix} \end{aligned}$$

We compute the elements of this covariance matrix by first computing the impulse responses of the $\frac{\partial a_t}{\partial \phi_k}$ and $\frac{\partial a_t}{\partial \theta_l}$. For example, to compute the covariance between:

$$w_t = \frac{\partial a_t}{\partial \phi_k} = \frac{-1}{\phi(q^{-1})} a_{t-k} \tag{5.45}$$

and

$$x_t = \frac{1}{\theta(q^{-1})} a_{t-l} \tag{5.46}$$

first compute the impulse response of each system:

$$\begin{aligned} w_t &= (1 + \zeta_1 B + \zeta_2 B^2 + \dots) a_{t-k} \\ x_t &= (1 + \xi_1 B + \xi_2 B^2 + \dots) a_{t-l} \end{aligned} \tag{5.47}$$

then,

$$\varrho_{xw}(u) = E\{x_t w_{t+u}\} = \sigma^2 \sum_{m=0}^{\infty} \zeta_{m+u} \xi_m \tag{5.48}$$

$$\varrho_x(u) = E\{x_t x_{t+u}\} = \sigma^2 \sum_{m=0}^{\infty} \xi_{m+u} \xi_m \tag{5.49}$$

Although the expressions (5.49) and (5.48) are not exact unless the infinite sum is computed in each case, in practice, the error incurred by truncating the sum at an appropriately high lag is negligible unless a series is virtually divergent.

When there exist common roots in the polynomials $\phi(B)$ and $\theta(B)$, the variance-covariance matrix for the parameters is singular. For cases where there are very similar roots in $\phi(B)$ and $\theta(B)$, the variance-covariance matrix will be ill conditioned, the parameter estimates will be highly correlated with each other, and the uncertainty in

the individual parameters will be large.

Evaluating $tr\left(\tilde{\Omega}_n{}^{-1}\Omega_n^*\right)$ in (5.28) involves computing the values of the elements of two $n \times n$ matrices, finding the inverse of one of these matrices, and then taking the product of that inverse matrix with the other matrix. For large n, this may involve a prohibitively large number of calculations. Therefore, there is a need for an efficient algorithm by which to compute $tr\left(\tilde{\Omega}_n{}^{-1}\Omega_n^*\right)$.

One approach is to exploit efficiency of Levinson recursion. Let

$$\boldsymbol{A} = \tilde{\Omega}_n{}^{-1}\Omega_n^* \tag{5.50}$$

Then, $\tilde{\Omega}_n \boldsymbol{A} = \Omega_n^*$. Because we are interested in the trace of $\boldsymbol{A}$, we need not compute all of the elements of $\boldsymbol{A}$ but only the diagonal elements. Let

$$\begin{aligned} \Omega_n^* \boldsymbol{a}_1 &= \boldsymbol{\omega}_1^* && (5.51) \\ &\vdots \quad \vdots && (5.52) \\ \Omega_n^* \boldsymbol{a}_n &= \boldsymbol{\omega}_n^* \end{aligned}$$

We need to solve for the first element of $\boldsymbol{a}_1$, the second element of $\boldsymbol{a}_2$, and so on. Using Cramer's rule,

$$\boldsymbol{a}_i(i) = \frac{|\Omega_n^{*(i)}|}{|\Omega_n^*|} \tag{5.53}$$

where $\boldsymbol{a}_i(i)$ is the i^{th} element of $\boldsymbol{a}_i$ and $\Omega_n^{*(i)}$ is the matrix Ω_n^* having its i^{th} column replaced by $\boldsymbol{\omega}_i^*$. Since Ω_n^* is a Toeplitz matrix, its determinant can be evaluated quickly using Levinson recursion.

Lam and Watts (1991) based their profiling calculations on an expression for the exact likelihood function of an ARMA model develop by Ansley (1979). However, many other expressions and algorithms for computing the exact likelihood have been

proposed, including those by Newbold (1974), Ali (1977), and Ljung and Box (1979). Harvey and Phillips (1979), and Åström (1980), among others, have developed algorithms based on the Kalman filter. Expressions for the exact likelihood function for vector ARMA(p,q) processes have been developed by Osborn (1977), Phadke and Kedem (1978), Hillmer and Tiao (1979), and Nicholls and Hall (1979). These, of course, can also be used for the special case of univariate ARMA models.

To be consistent with the work of Lam and Watts (1991), we have based our calculations on the transformation of Ansley. The expressions for $\tilde{\mathbf{\Omega}}_n$ and $\mathbf{\Omega}_n^*$ are developed below.

Ansley's algorithm is based on transforming a time series as follows:

$$z_t = \begin{cases} y_t, & t = 1, \ldots, m \\ \phi(B)y_t, & t = m+1, \ldots, n \end{cases} \tag{5.54}$$

where $m = max(p, q)$. Let

$$v_t = \theta(B)a_t = \phi(B)y_t \tag{5.55}$$

The series v_t is autocorrelated only up to lag q. Then, the covariance matrix for z_t has a maximum bandwidth of m for the first m rows and a bandwidth of q thereafter (Ansley, 1979), where the bandwidth is the number of nonzero elements in the row. Let $\gamma_y(i)$ be the autocovariance of y_t at lag i, $\gamma_v(i)$ be the autocovariance of v_t at lag

i, and $\gamma_{yv}(i)$ be the cross-covariance of y_t and v_t at lag i. Then,

$$\boldsymbol{\Omega}_{n,z} = \begin{bmatrix} \gamma_y(0) & \cdots & \cdots & \gamma_y(m) & & & & \\ \vdots & \ddots & & & \gamma_{vy}(q) & & 0 & \\ \vdots & & & \vdots & \vdots & \ddots & & \\ \gamma_y(m) & \cdots & \cdots & \gamma_y(0) & \gamma_{vy}(1) & \cdots & \gamma_{vy}(q) & \\ & \gamma_{vy}(q) & \cdots & \gamma_{vy}(1) & \gamma_v(0) & \ddots & \gamma_v(q) & \\ & & \ddots & \vdots & \vdots & \ddots & \ddots & \\ & & & \gamma_{vy}(q) & \vdots & \ddots & \ddots & \gamma_v(q) \\ & 0 & & & \gamma_v(q) & \ddots & \ddots & \vdots \\ & & & & & \gamma_v(q) & \cdots & \gamma_v(0) \end{bmatrix}_{n \times n} \tag{5.56}$$

We compute $\tilde{\boldsymbol{\Omega}}_n$ and $\boldsymbol{\Omega}_n^*$ on the basis of the transformed series $\tilde{z}_t$ and z_t^*, respectively. These series are defined below. Let $\tilde{\boldsymbol{\Omega}}_{n,z}$ and $\boldsymbol{\Omega}_{n,z}^*$ be the variance-covariance matrices for the transformed series. When computing $\tilde{\boldsymbol{\Omega}}_{n,z}$ we assume that the model for y_t is:

$$\bar{\phi}(B) y_t = \tilde{\theta} a_t \tag{5.57}$$

Then, the transformed series is:

$$\tilde{z}_t = \begin{cases} y_t, & t = 1, \ldots, m \\ \tilde{\phi}(B) y_t, & t = m+1, \ldots, n \end{cases} \tag{5.58}$$

$$= \begin{cases} \frac{\tilde{\theta}(B)}{\tilde{\phi}(B)} a_t, & t = 1, \ldots, m \\ \tilde{\theta}(B) a_t, & t = m+1, \ldots, n \end{cases} \tag{5.59}$$

and $\tilde{\Omega}_{n,z}$ is a banded matrix of the form shown in (5.56).

The variance-covariance matrix $\Omega^*_{n,z}$ is equal to $E\{\boldsymbol{z}_n^T \boldsymbol{z}_n\}$. When data are used to compute profile t plots the observed variance-covariance matrix is computed. In that case, the values of $\boldsymbol{z}_n$ would be values of y_t generated by the true process transformed using $\tilde{\phi}(B)$. Therefore, when computing $\Omega^*_{n,z}$ for expected profiling we employ the true process model

$$\phi^*(B) y_t = \theta^* a_t \tag{5.60}$$

Then, we apply the same transformation as used to compute $\tilde{\Omega}_{n,z}$. That is to say we compute $\Omega^*_{n,z}$ based on the transformed series:

$$z_t^* = \begin{cases} y_t, & t = 1, \ldots, m \\ \tilde{\phi}(B) y_t, & t = m+1, \ldots, n \end{cases} \tag{5.61}$$

$$= \begin{cases} \frac{\theta^*(B)}{\phi^*(B)} a_t, & t = 1, \ldots, m \\ \frac{\theta^*(B)}{\phi^*(B)} \tilde{\phi}(B) a_t, & t = m+1, \ldots, n \end{cases} \tag{5.62}$$

Define:

$$\tilde{v}_t = \frac{\theta^*(B)}{\phi^*(B)} \tilde{\phi}(B) a_t \tag{5.63}$$

Whereas the polynomial operators $\hat{\phi}(B)$ and $\tilde{\phi}(B)$ cancel in the case of v_t, they do

not cancel in (5.63). The expression for $\boldsymbol{\Omega}^*_{n,z}$ is:

$$\boldsymbol{\Omega}^*_{n,z} = \begin{bmatrix} \gamma^*_y(0) & \cdots & \gamma^*_y(m) & \gamma_{\tilde{v}_y}(m+1) & \cdots & \gamma_{\tilde{v}_y}(n) \\ \vdots & & \vdots & \vdots & \ddots & \\ \gamma^*_y(m) & \cdots & \gamma^*_y(0) & \gamma_{\tilde{v}_y}(1) & \cdots & \gamma_{\tilde{v}_y}(n-m) \\ \gamma_{\tilde{v}_y}(m+1) & \cdots & \gamma_{\tilde{v}_y}(1) & \gamma_{\tilde{v}}(0) & \cdots & \gamma_{\tilde{v}}(n-m+1) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{\tilde{v}_y}(n) & \cdots & \gamma_{\tilde{v}_y}(n-m) & \gamma_{\tilde{v}}(n-m+1) & \cdots & \gamma_{\tilde{v}}(0) \end{bmatrix}_{n \times n} \tag{5.64}$$

Note that while $\tilde{\boldsymbol{\Omega}}_{n,z}$ is banded, in general, $\boldsymbol{\Omega}^*_{n,z}$ is not. Also, $\tilde{\boldsymbol{\Omega}}^{-1}_{n,z}$ is not, in general, a banded matrix; however, Ma (1997) has proposed an expression for the efficient computation of this inverse.

As previously noted, finding values of the exact likelihood function or the expected value of the exact likelihood function can be computationally intensive. For large n, it may be prohibitively expensive to do the number of calculations required to solve the sequence of optimization problems involved in profiling. The expression (Box and Jenkins, 1976)

$$L(\boldsymbol{\theta}|\boldsymbol{y_n}) = (2\pi\sigma_a^2)^{-n/2} \exp\left(\frac{-\boldsymbol{a}^T\boldsymbol{a}}{2\sigma_a^2}\right) \tag{5.65}$$

is exact. The variance-covariance matrix $\boldsymbol{\Omega}_n$ disappears because of the Jacobian of the transformation from $\boldsymbol{y_n}$ to $\boldsymbol{a}$. However, to compute $\boldsymbol{a}$, it is often necessary to make assumptions about the initial conditions of the system. Then, in practice, the values of the likelihood function computed based on (5.65) are approximate in that they are conditional upon the assumptions about the intial conditions of the

system. Evaluation of (5.65) requires a significantly smaller computational effort than evaluation of the exact likelihood given in (5.14). Although all expected profiles and n-plots shown in this paper were computed on the basis of the exact likelihood function, we recommend the use of the conditional likelihood function when appropriate.

5.5 Illustrative Examples

Two data sets from the literature are used to illustrate the concepts developed in the preceding sections. Tabulated information about the data sets, the fitted models and the inference results are given in Section 5.10.

For Example 1, an ARIMA(2,0,2) model was fitted to the mean centered "Housing Permits" data used by Pankratz (1983). The profile t plots for the estimated parameters are shown in Figure 5.1. The corresponding expected profiles are shown in Figure 5.2. The shapes of the expected profiles are remarkably similar to those of the profile t plots based on the observed data. The expected profile plots reliably capture the information about the degree of nonlinearity to be expected. However, notice that the likelihood intervals based on the profile t plots are wider than the corresponding expected likelihood intervals. This is a natural consequence of that fact that peculiarities (eg. outliers) in a measured data set beyond the behavior dictated by the model, may inflate the uncertainties in the estimates of the parameters in a proposed model; since expected profiles are computed in the absence of data, they will not reflect such contributions to the uncertainty.

Example 2 is based on an ARIMA(2,0,3) model fitted to the "Coal Production" data used by Pankratz (1983). Profile t plots for the estimated parameters are shown in Figure 5.3. The corresponding expected profiles are shown in Figure 5.4. The location of the MLE estimates of the autoregressive parameters of Example 2 relative to the stability boundaries is shown in Figure 5.6. Example 2 is particularly inter-

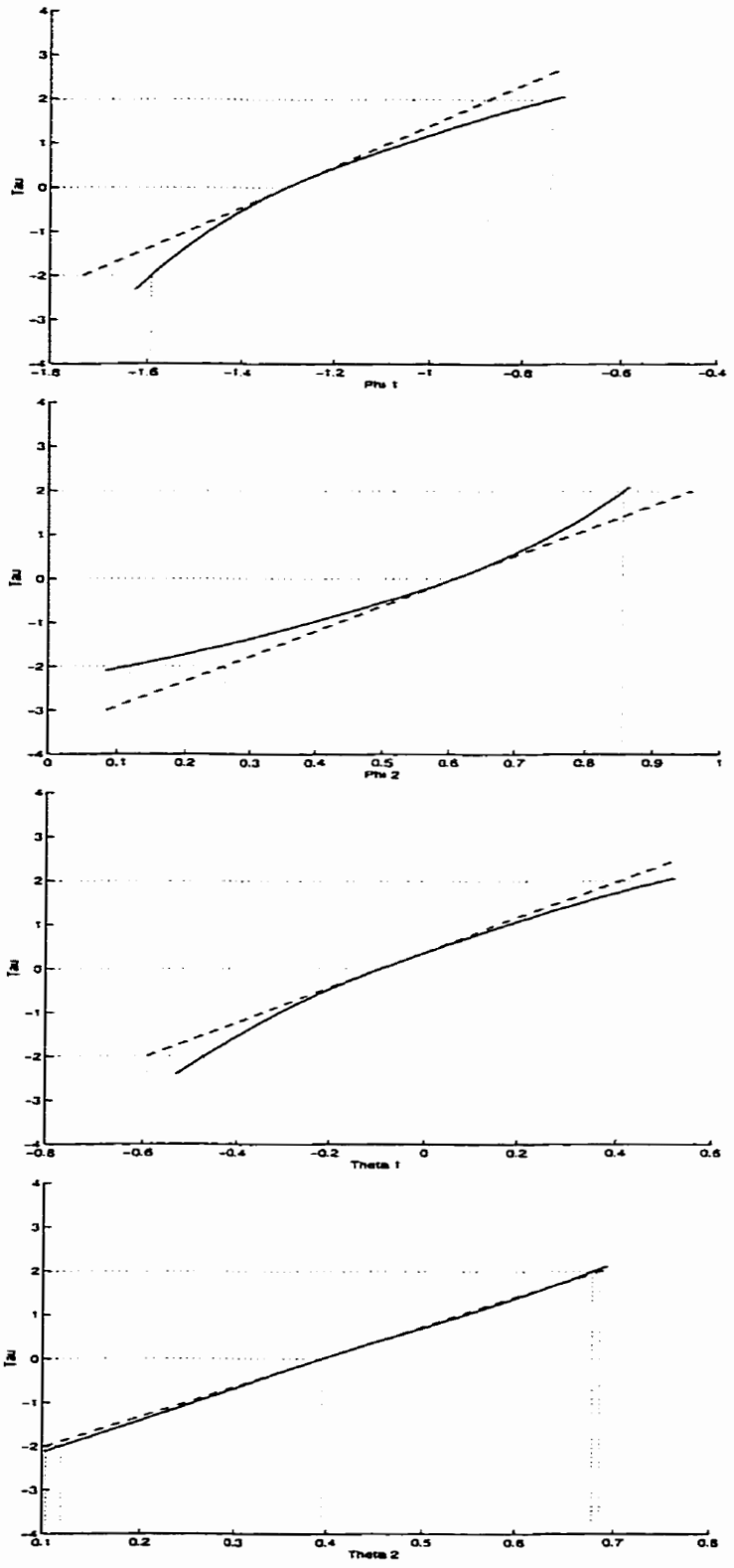

Figure 5.1: Profile t plots for the parameters of Example 1. - - - reference line; — Profile t plot; ··· maximum likelihood estimates and limits of the 95 % linearization and likelihood intervals.

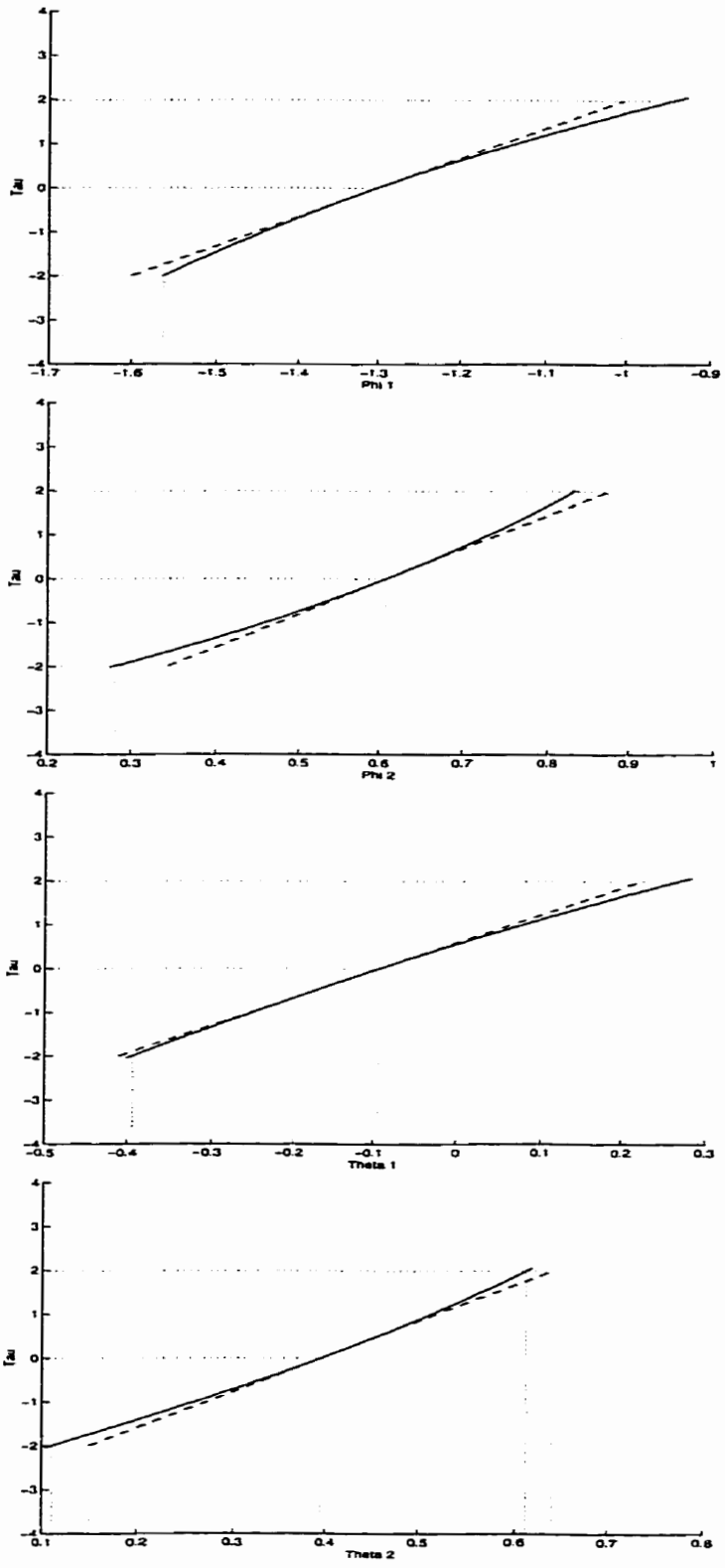

Figure 5.2: Expected profile t plots for the parameters of Example 1. - - - reference line; — Profile t plot; $\cdots$ maximum likelihood estimates and limits of the 95 % linearization and likelihood intervals.

esting in that the peculiar shapes of the profile t plots for the parameters ϕ_1 and θ_1 are well approximated by the corresponding expected profiles, although the same is not true for the other three parameters. Overall, the expected profiles appropriately warn of the large uncertainties and nonlinearities associated with the estimates of the parameters. The "cliff" which is a feature of the profile t curve for ϕ_2 is typical of the behaviour of these plots when the vector of parameter values comes very close to a stability/invertibility boundary.

The profile t plot for parameter ϕ_2 differs most noticeably from the expected profile. A small scale simulation study was done to determine whether this was simply the result of an anomaly in the data or whether the expected profile was failing to identify an important nonlinearity of the model. Ten data sets, each of 96 observations (since there were 96 observations in the observed data set), were generated based on the fitted model for Example 2 (i.e., based on the maximum likelihood estimates for the parameters given in Appendix A). Three of the ten data sets resulted in profile t plots for ϕ_2 which corresponded well to the expected profile. The other seven data sets resulted in profile t plots for ϕ_2 which resembled that based on the original data set. Therefore, we concluded that the lack of agreement between the profile t plot for ϕ_2 for the original data and the corresponding expected profile is a manifestation of a peculiarity in the data generated by this process. Anomalies in data can have a dramatic affect on both the maximum likelihood estimates of parameters and the uncertainties in the estimates. Clearly, expected profiling can never inform about data-related anomalies. However, in most cases that we have examined, the expected profiles have been successful in predicting the major features observed in the corresponding profile t plots.

The profile t plots and the expected profile t plots can assume a wide range of shapes. That is, the degree of nonlinearity can vary dramatically from parameter to parameter and from example to example. The profile t plots for all of the parameters

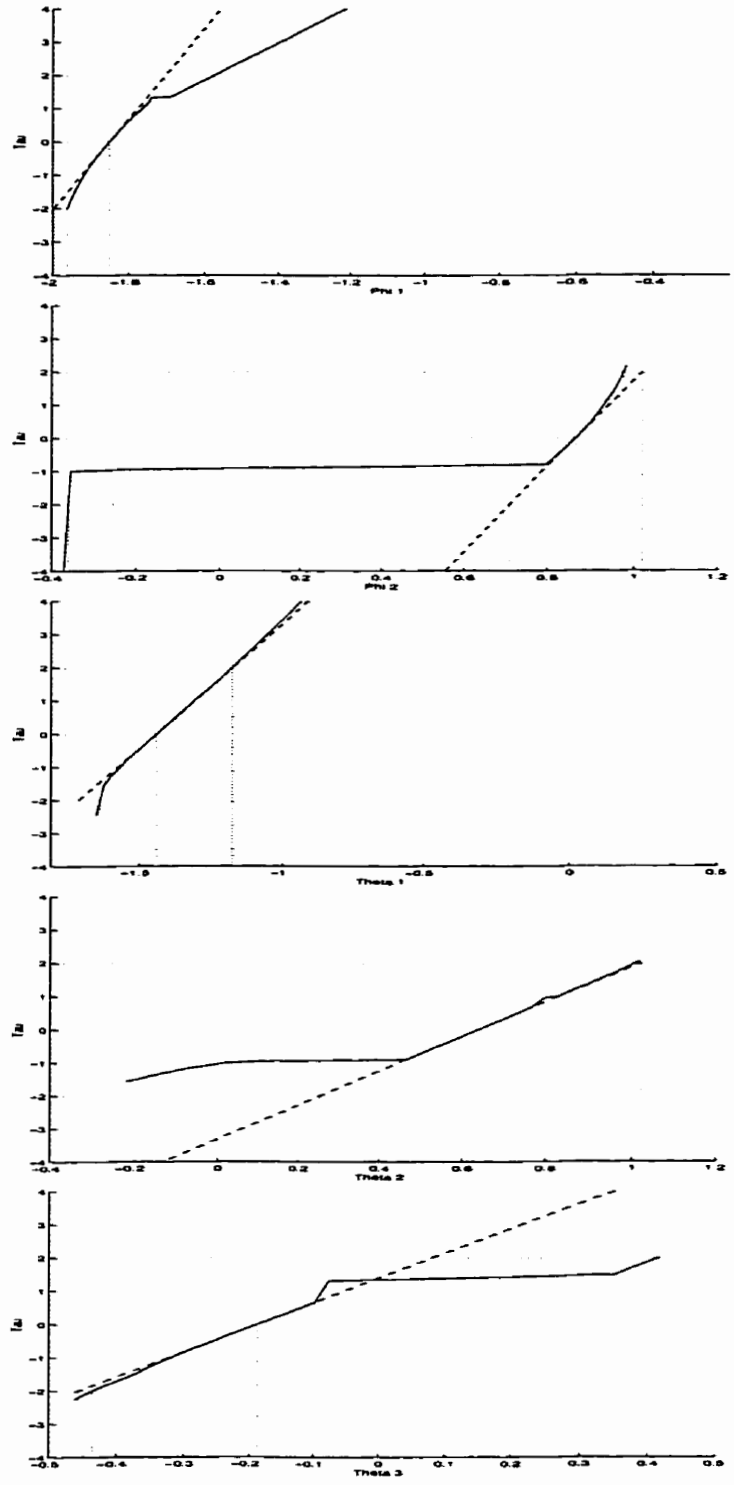

Figure 5.3: Profile t plots for the parameters of Example 2. - - - reference line; — Profile t plot; ··· maximum likelihood estimates and limits of the 95 % linearization and likelihood intervals.

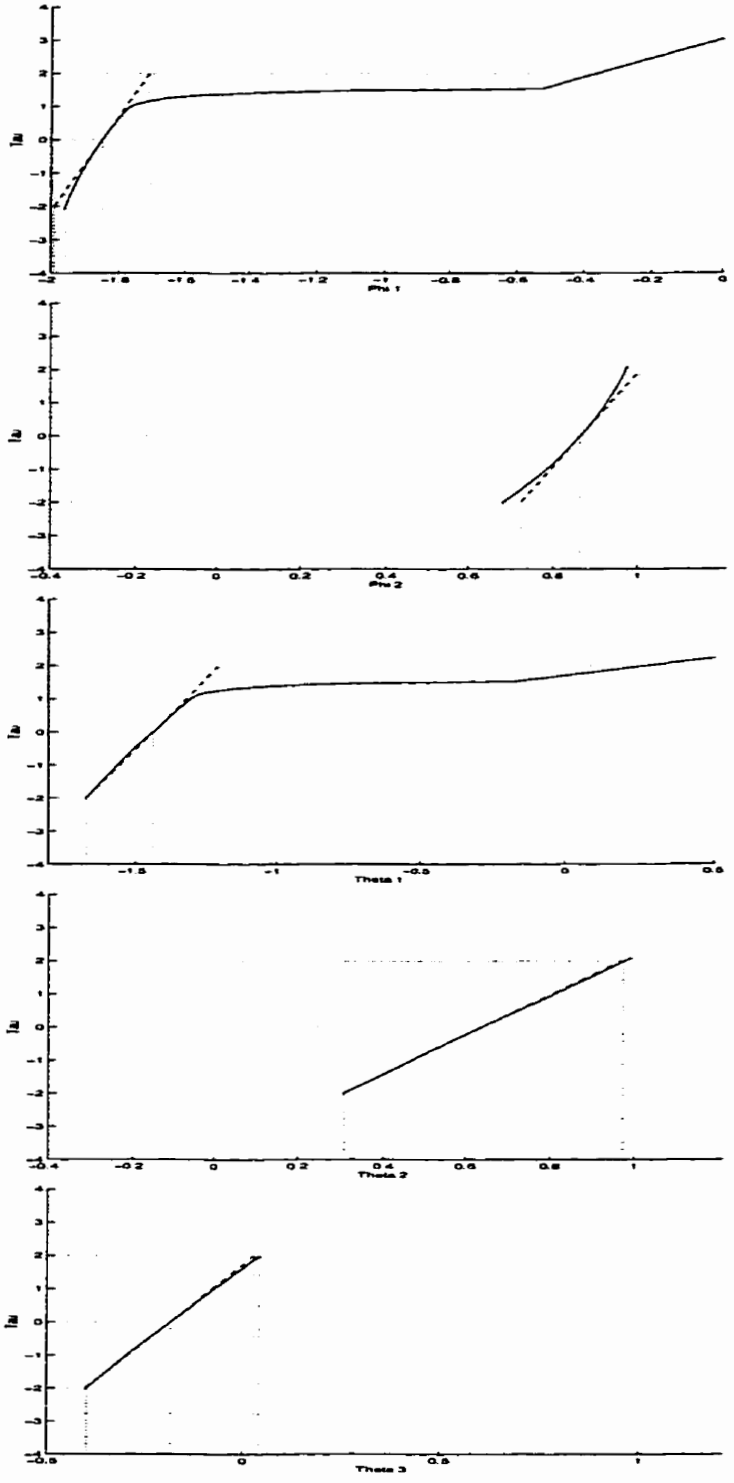

Figure 5.4: Expected profile t plots for the parameters of Example 2. - - - reference line; — expected profile; $\cdots$ maximum likelihood estimates and limits of the 95 % linearization and likelihood intervals.

of Example 1, as well as the expected profiles for these parameters, behave approximately linearly, and likelihood intervals for these parameters are well approximated by linearization confidence intervals. In contrast, the profile t plots for some parameters of Example 2 are drastically nonlinear, and linearization confidence intervals for these parameters are misleading.

The nonlinearities displayed in Examples 1 and 2 are consistent with the work of Lam and Watts (1991), which showed that profile t plots for parameters in time series models tend to behave nonlinearly when the parameters are close to a stability/invertibility boundary. The estimates of the parameters for the model in Example 1 are well within the stability region (see Figure 5.5), and the profile t plots are relatively linear, whereas the *vector* of parameter estimates in Example 2 is close to a stability/invertibility boundary and the profile t plots are drastically nonlinear. Note that it is not necessarily the proximity of individual parameters to their respective individual stability/invertibility limits, but rather the proximity of the vector of AR parameters in p-space to the closest joint stability boundary, and/or the proximity of the q-dimensional vector of MA parameters to the nearest invertibility boundary, that is the determining factor. This point is made clear in Example 2, where none of the five parameters is "very" close to its *individual* stability/invertibility limits, but the *vector* of AR parameters in 2-dimensional space is close to a stability limit (see Figure 5.6). This proximity causes several of the parameters to display nonlinearity.

Expected profiles can be used to obtain reasonable estimates of likelihood intervals for the parameters of a model *a priori* to data collection. Figure 5.7 shows how the limits of the likelihood intervals for the parameters of Example 2 are expected to change as n increases. At $n = 100$, the likelihood limits are very wide, indicating that a high degree of uncertainty in the parameter estimates should be expected if the estimates are based on less than 100 observations. Indeed, the estimates based on the 96 observations of coal production were highly uncertain. As n increases, the intervals

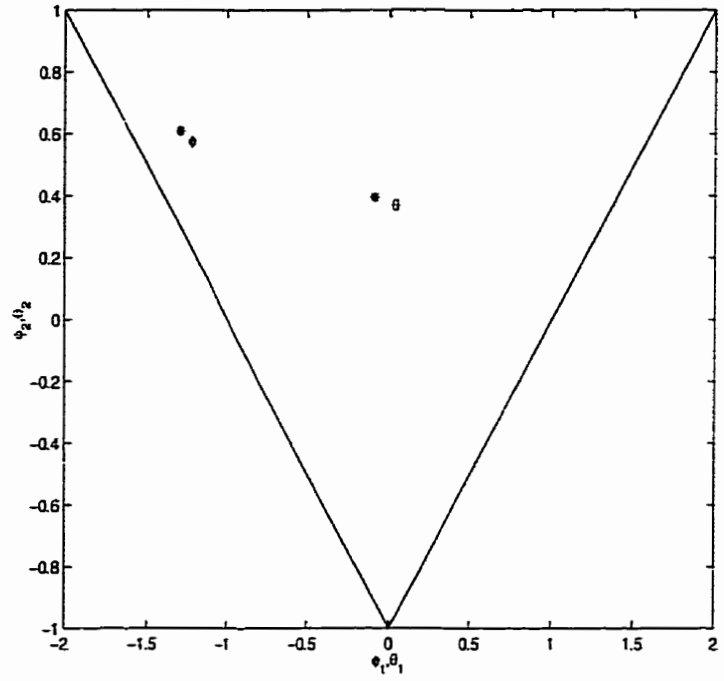

Figure 5.5: The location of the ML estimates of the AR and MA parameters of Example 1 relative to the stability boundaries. Key: – stability/invertibility boundary; $*$ ϕ vector of ML estimates of the $\phi(B)$ parameters; $*$ θ vector of ML estimates of the $\theta(B)$ parameters.

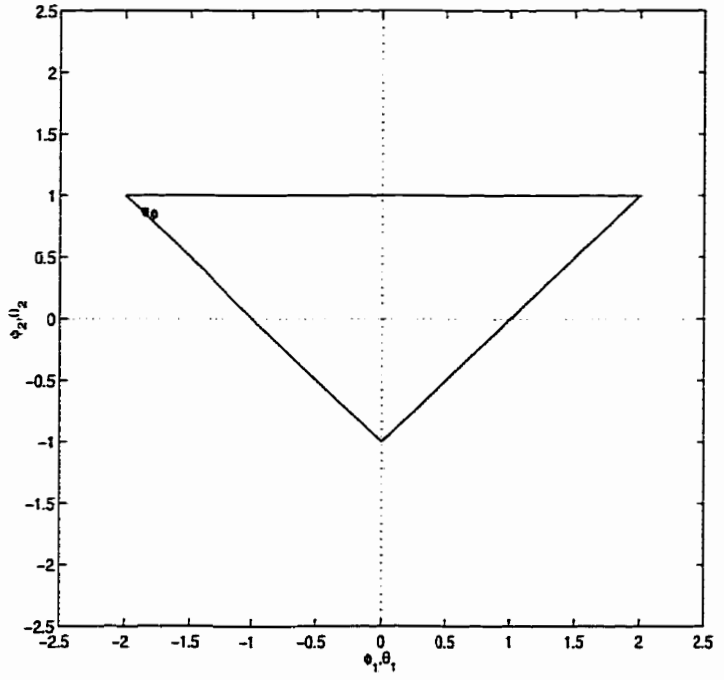

Figure 5.6: The location of the ML estimates of the autoregressive parameters of Example 2 relative to the stability boundaries.

become narrower, indicating that less uncertainty can be expected in the parameter estimates. This is a reflection of the increase in the amount of information available about the process. The dashed lines on these plots represent the linear approximation intervals (i.e., the intervals based on the Cramer Rao lower bounds). The value of n at which the expected likelihood intervals coincide with the linearization intervals provides an indication of how much data would be required before asymptotic results would be appropriate. The n-plots for Example 2 show that for $n < 200$, the linearization confidence intervals would be very misleading.

Figure 5.8 shows the expected profiles for the parameters at $n = 100$ and $n = 500$. As anticipated, the expected profiles based on 500 observations behave more linearly than those based on 100 observations, consistent with the asymptotic theory which states that as n approaches infinity, the distribution of the uncertainty in the parameter estimates approaches a spherical normal distribution. Note that the expected profiles in these figures are plots of $\tau(g)$ versus $\delta(g)$. Plots of n versus the likelihood limits also convey information about the increase in linearity of the estimates as n increases, as the likelihood intervals approach the linearization confidence intervals with increasing n.

Plots of n versus $g(\boldsymbol{\theta})$ provide information to experimenters about the relative value of acquiring additional data from a process. For example, Figure 5.8 indicates that for this ARIMA(2,0,3) model, considerable improvement in the quality of the parameter estimates would be expected by increasing the number of observations from 100 to 200. However, the gain from increasing the number of observations from 400 to 500 would be minimal. Thus, expected profiling is a tool which can be used to judge the cost/benefit ratio of increasing the length of a data set.

The comparisons between the expected profiles and the profile t for these two examples were done knowing the MLEs of the parameters. However, we propose expected profiling as a means for making judgments about acceptable lengths of data

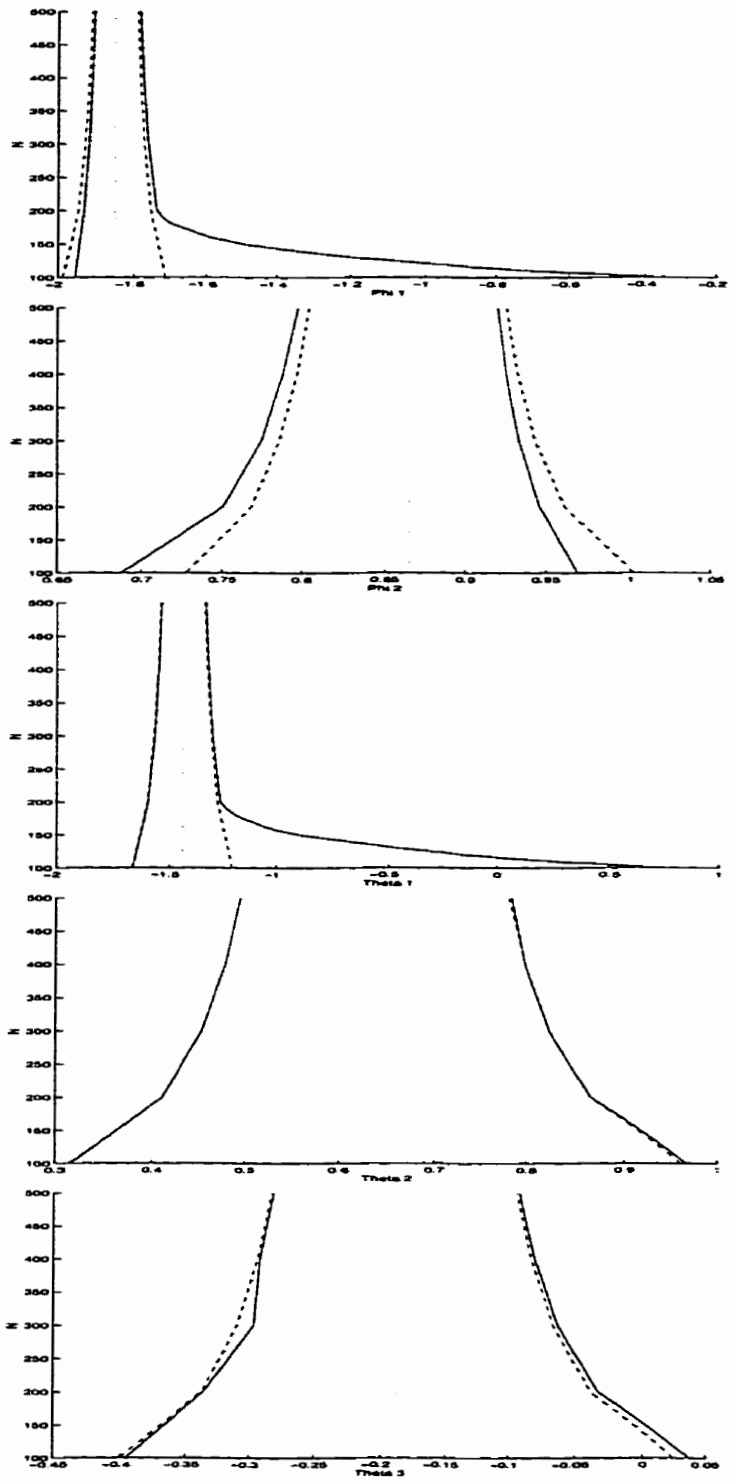

Figure 5.7: Plot of n versus the uncertainty limits for the parameters of Example 2. - - - 95 % linearization intervals; — 95 % likelihood limits from expected profiles; ··· maximum likelihood estimate.

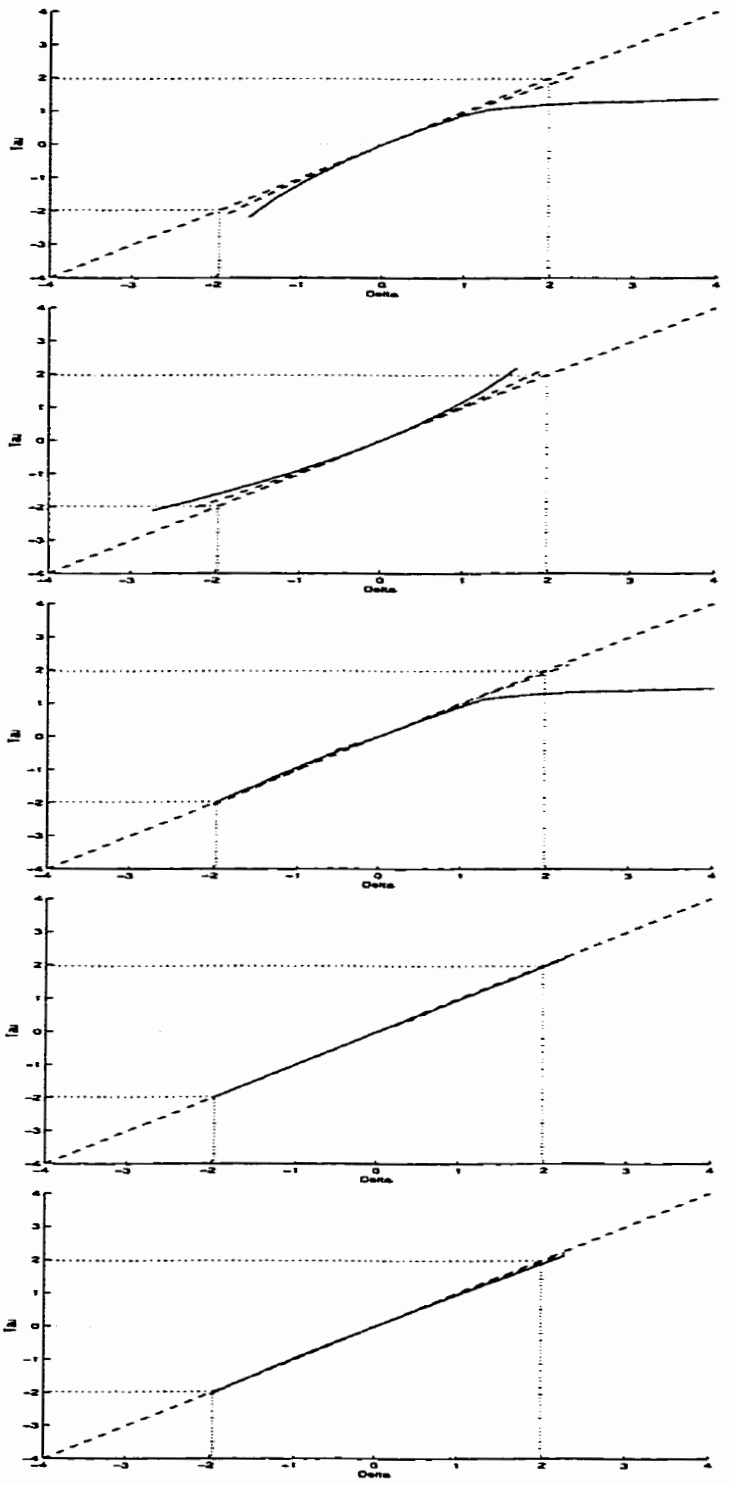

Figure 5.8: Expected profile plots for the parameters of Example 2. - - - reference line; - - - expected profile for $n = 500$; — expected profile for $n = 100$; $\cdots$ maximum likelihood estimates and limits of the 95 % linearization and likelihood intervals.

sets prior to experimentation. This requires that guesses be made about the values of the parameters and the form of the model. It is important to appreciate how sensitive the expected profiles are to the values of the parameters. It has been our experience that the shapes of the expected profiles and n-plots are insensitive to small changes in the values of the parameters. The shape depends primarily on how close the vector of parameters is to a stability or invertibility boundary. Typically, a data analyst will have a good idea about whether a time series is close to being non-stationary, and can thereby make a reasonably reliable *a priori* judgement about how close the parameter vector is to a stability boundary.

Because expected profile plots are relatively insensitive to the values of the parameters, we propose that a library of expected profile plots for common ARMA models, over a range of parameter values, be established. This would serve as a quick reference for data analysts and would save the considerable computational effort required to generate the n-plots based on the exact likelihood function.

5.5.1 Full Likelihood Estimation versus Conditional Likelihood Estimation

As discussed previously in Section 5.4.1, finding values of the exact likelihood function or the expected value of the exact likelihood function can be computationally intensive. For large n, it may be prohibitively expensive to do the number of such calculations required to solve the sequence of optimization problems involved in profiling. The expression given in (5.65) requires a significantly smaller computational effort than evaluation of the exact likelihood given in (5.14). However, the expected profiles obtained based on (5.65) may be significantly different than those based on the full likelihood function if the vector of parameter estimates is close to a stability/invertibility boundary.

Figures 5.9 and 5.10 show the expected profiles for ϕ_1 based on both the full and

approximate likelihood functions for Examples 1 and 2, respectively. The computational times required to generate these expected profiles are listed in Table 5.1. For Example 1, the profiles based on the full and approximate likelihood functions are almost indistinguishable from each other; however, there was an order of magnitude difference in the computational time required to produce the profiles. For this example, because the vector of parameter values is far from all stability/invertibility boundaries, we recommend the use of the approximate likelihood function. For Example 2 where the vector of parameter values is close to a stability/invertibility boundary, the expected profile based on the approximate likelihood function differs significantly from that based on the full likelihood function. In this case, the extra computational time should be expended to obtain expected profiles which are reliable. To appreciate the magnitude of the difference in computational effort between the full likelihood and the approximate likelihood approaches, consider that the time to compute the n-plot for Example 1 based on the approximate likelihood function was 2.6 hours, while the time to compute the same based on the expression for the exact likelihood was 173 hours. Clearly, work is still required to make the algorithm more efficient in terms of computation time so that it can be used routinely.

Table 5.1: Comparison of the Computational Times Required for Expected Profiling Based on the Full and Approximate Expressions for the Likelihood Function

	Time using Approx. Likelihood (seconds)	Time using Approx. Likelihood (seconds)
Example 1	1.612×10^3	1.144×10^4
Example 2	5.117×10^3	4.134×10^4

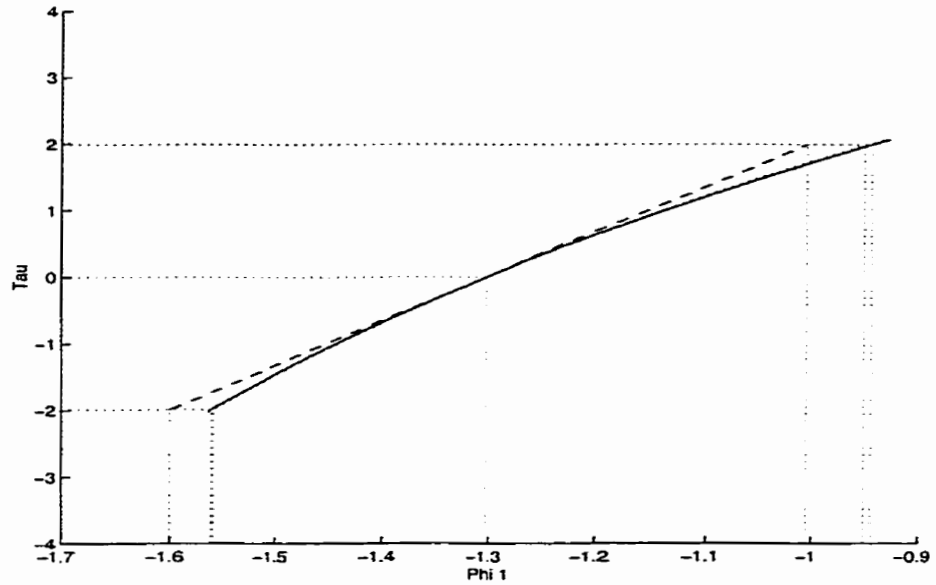

Figure 5.9: The expected profiles for ϕ_1 based on both the full and approximate likelihood functions for Examples 1 with $n = 84$. Key: - - reference line; $\cdots$ expected profile based on the approximate likelihood function, – expected profile based on the full likelihood function.

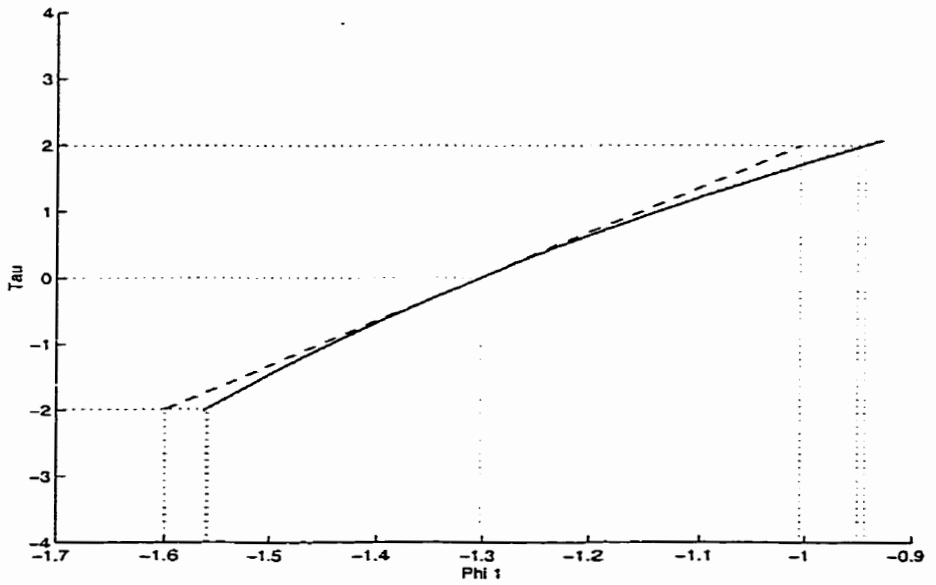

Figure 5.10: The expected profiles for ϕ_1 based on both the full and approximate likelihood functions for Examples 2 with $n = 96$. Key: - - reference line; $\cdots$ expected profile based on the approximate likelihood function, – expected profile based on the full likelihood function.

5.6 The Delta Parameterization

The previous sections focussed on using expected profiling to investigate the expected behavior of estimates of individual parameters of ARMA models. In this section we give an example of using expected profiling to study a function of parameters $g(\boldsymbol{\theta})$. Middleton and Goodwin (1990) have proposed the use of the δ operator (not to be confused with the function $\delta(g)$) instead of the backshift operator B. They have found that time series models and discrete transfer function models are more naturally expressed using this operator since the link to the continuous time representation is more obvious, and the numerical properties of calculations based on the δ operator are superior to those using the backshift operator (Middleton and Goodwin, 1986). The δ operator is defined by:

$$\delta = \frac{z-1}{\Delta} \tag{5.66}$$

so that

$$z = \Delta\delta + 1 \tag{5.67}$$

where Δ is the sampling period, and z is the forward shift operator. To express an ARMA model in terms of δ, we simply make the substitution given by (5.67). Then, we express the new model as:

$$(\phi_1^d\delta^p + \phi_2^d\delta^{p-1} + \ldots + 1)y_t = (\theta_1^d\delta^q + \theta_2^d\delta^{q-1} \ldots + 1)(b)a_t \tag{5.68}$$

where ϕ_i^d and θ_i^d are the parameters of the δ model. Equivalently,

$$(\phi_1^d\delta^p + \phi_2^d\delta^{p-1} + \ldots + 1)y_t = (\theta_1^d\delta^q + \theta_2^d\delta^{q-1} \ldots + 1)a_t^* \tag{5.69}$$

where $b = \frac{1+\theta_1+\ldots+\theta_q}{1+\phi_1+\ldots+\phi_p}$, $a_t^* = ba_t$, and $\sigma_a^{2^*} = b^2\sigma_a^2$. For the time series models discussed in this work, $\Delta = 1$.

The new vector of parameters $\boldsymbol{\theta}^d$ can be shown to be a function of the parameters of the original model. For example, the model:

$$(1 + \phi_1 B + \phi_2 B^2)y_t = (1 + \theta_1 B + \theta_2 B^2)a_t \tag{5.70}$$

may be written equivalently as:

$$(z^2 + \phi_1 z + \phi_2)y_t = (z^2 + \theta_1 z + \theta_2)a_t \tag{5.71}$$

Substituting (5.67) into (5.71),

$$\begin{aligned} \left(\frac{\Delta^2}{1+\phi_1+\phi_2}\delta^2 + \frac{\Delta(\phi_1+2)}{1+\phi_1+\phi_2}\delta + 1\right) y_t &= \\ \left(\frac{\Delta^2}{1+\theta_1+\theta_2}\delta^2 + \frac{\Delta(\theta_1+2)}{1+\theta_1+\theta_2}\delta + 1\right)\left(\frac{1+\theta_1+\theta_2}{1+\phi_1+\phi_2}\right) a_t & \end{aligned} \tag{5.72}$$

Therefore, with $\Delta = 1$,

$$\phi_1^d = \frac{1}{1+\phi_1+\phi_2} \tag{5.73}$$

and so on.

We wish to compare the behavior of the estimates of the parameters of the delta model with that of the estimates of the parameters of an equivalent ARMA model expressed using the traditional backshift operator. Figure 5.11 shows the expected profile plots for the $\boldsymbol{\theta}^d$ parameters of the model of Example 1 expressed in terms of the δ operator. The expected profile plots for the δ parameters are much more nonlinear than the expected profile plots for the original parameters. The expected profiles suggest that the values of the parameters in the delta model can not be negative.

Indeed this is the case because for stability, $1 + \phi_1 + \phi_2$ and $1 + \theta_1 + \theta_2$ must be greater than 0.25, and $\phi_1 + 2$ and $\theta_1 + 2$ must be greater than 1. The locations of the vectors of parameter estimates for the $\phi^d(\delta)$ and $\theta^d(\delta)$ parameters are shown in Figure 5.12. Whereas the stability/invertibility region is a closed triangle in the space of the original parameters, in the space of the delta-model parameters, the stability/invertibility region is infinite with boundaries defined by:

$$\phi_2^d = 2\phi_1^d + 0.5 \tag{5.74}$$

$$\phi_2^d = 1 \tag{5.75}$$

$$0.25 \leq \phi_1^d \leq \infty \tag{5.76}$$

$$1 \leq \phi_2^d \leq \infty \tag{5.77}$$

The vector of $\phi^d(\delta)$ parameters is closer to a stability/invertibility boundary than is the vector of AR parameters for the model expressed in terms of B. Therefore, the parameters of the model expressed in terms of the δ operator display significantly more nonlinearity.

5.7 Expected Profiling and Nonlinear Regression Models

We have focussed on expected profiling of functions of parameters in ARMA time series models. However, expected profiling can also be used in the context of other classes of models including nonlinear regression models.

Recall that expected profiling is based on the expression

$$E\{\tau^2\} = tr\left(\tilde{\Omega}_n{}^{-1}\Omega_n^*\right) - n + \ln\left(\frac{\left|\tilde{\Omega}_n\right|}{|\Omega_n^*|}\right)$$

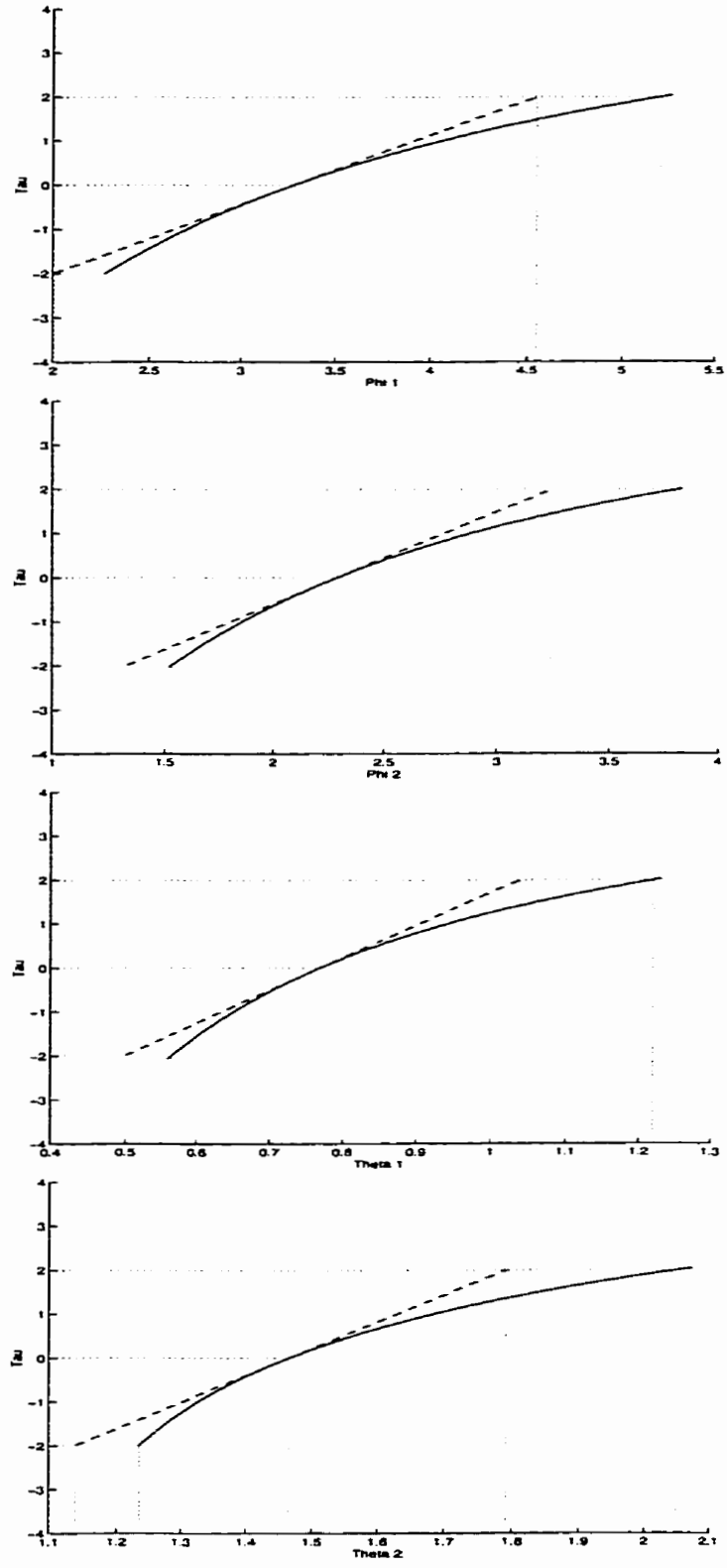

Figure 5.11: Expected profile t plots for the parameters of the "delta model" of Example 1. - - - reference line; — expected profile; ··· maximum likelihood estimates and limits of the 95 % linearization and likelihood intervals.

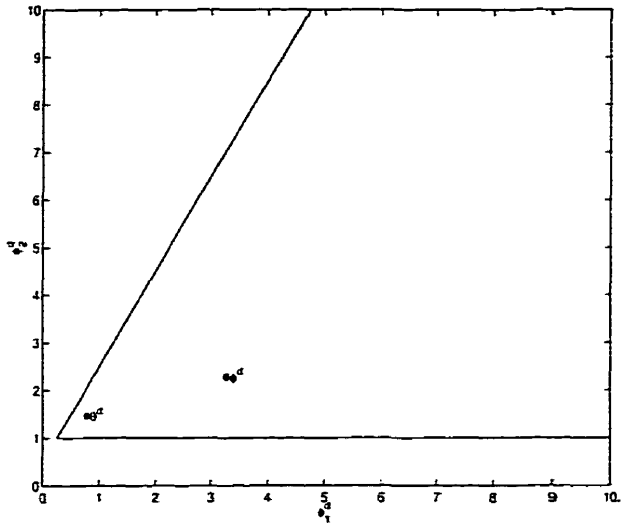

Figure 5.12: The location of the ML estimates of the parameters of the model for Example 1 expressed in terms of the δ operator, relative to the stability boundaries. Key: – stability/invertibility boundary; $*$ ϕ vector of ML estimates of the $\phi(B)$ parameters; $*$ θ vector of ML estimates of the $\theta(B)$ parameters.

Therefore, expected profiling can be used with any class of model so long as the variance-covariance matrices $\tilde{\boldsymbol{\Omega}}_n$ and $\boldsymbol{\Omega}_n^*$ can be computed. For nonlinear regression models of the form:

$$y = f(\boldsymbol{x}, \boldsymbol{\theta}) + \epsilon \tag{5.78}$$

where ϵ is independently and identically distributed additive random error with variance σ^2. When profiling, we presume that the set of parameter values, $\tilde{\boldsymbol{\theta}}$ being considered at any iteration of the algorithm represents the vector of true values of the parameters, and that the proposed form of the model can adequately describe the true process. Under this assumption, $\tilde{\boldsymbol{\Omega}}_n = \boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. The elements of the the $\boldsymbol{\Omega}_n^*$ matrix under the assumption that the $\tilde{\boldsymbol{\theta}}$ represents the true values of the parameters are

$$\{\boldsymbol{\Omega}_n^*\}_{i,j} = \frac{1}{\sigma^2} cov(\boldsymbol{y}_i, \boldsymbol{y}_j) = E[\boldsymbol{y}_i - E\boldsymbol{y}_i][\boldsymbol{y}_j - E\boldsymbol{y}_j] \tag{5.79}$$

$$= \frac{1}{\sigma^2}[f(\boldsymbol{x}_i, \boldsymbol{\theta}^*) f(\boldsymbol{x}_j, \boldsymbol{\theta}^*) - f(\boldsymbol{x}_i, \boldsymbol{\theta}^*) f(\boldsymbol{x}_j, \tilde{\boldsymbol{\theta}}) - \tag{5.80}$$

$$f(\boldsymbol{x}_i, \bar{\boldsymbol{\theta}}) f(\boldsymbol{x}_j, \boldsymbol{\theta}^*) + f(\boldsymbol{x}_i, \bar{\boldsymbol{\theta}}) f(\boldsymbol{x}_j, \bar{\boldsymbol{\theta}}) + \sigma^2]$$

Note that to use expected profiling in the context of regression models, the complete experimental design must be known. That is, the set of conditions at which observations of y are to be taken must be specified *a priori*. In the case of expected profiling of a time series model, only the amount of data, n must be specified.

5.8 Conclusions

Expected profiles can serve as useful approximations to the profile t plots obtained from a realization of a process. They can capture the significant nonlinearity inherent in parameter estimates for models whose parameter vector is close to a stability/invertibility boundary. n-plots provide useful information about how much data may be required to obtain acceptable estimates of the parameters (or functions of parameters) of a model.

The theory of expected profiling has been developed and illustrated here using two examples. More work is needed to study how well these methods perform over a large number of data sets. Several useful ways of plotting the information obtained from the expected profiling methodology are proposed, and are intended to emphasize the utility of expected profiling in designing experiments. Especially when one has control over how much data can be collected, expected profiling can be an important tool for deciding how much data will be "enough". Even for cases where there is little control over data collection, expected profiling may be used as a means of deciding which approaches to computing inference results are appropriate, and for making qualitative judgments about the sources of any observed nonlinearities. Bates and Watts (1980), along with others, developed measures of nonlinearity to separate nonlinearity due to parameterization from nonlinearity due to the form of the model. Expected profiling may be used as a tool to separate nonlinearity due to the data itself from nonlinearity

due to the model and its parameterization.

We exploit the dependence of $E\{\tau^2\}$ on n to develop a methodology for use in designing the length of an experiment to collect time series data. We have also used expected profiling to address the issue of how much data is required for asymptotic results to apply. The n-plot shows how the likelihood limits for $g(\boldsymbol{\theta})$ are expected to change as the value of n changes. On the same graph, the expected confidence intervals based on the Cramer Rao lower bounds are also shown. The distance between the limits of the intervals based on expected profiling and those based on the Cramer Rao lower bounds provides a qualitative measure of how reliable the asymptotic results are likely to be for a given value of n.

5.9 Acknowledgements

The authors gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada and the School of Graduate Studies of Queen's University.

5.10 Appendix

Table 5.2: Data and Models Used in the Examples

	Example 1	Example 2
Source	Pankratz (1983)	Pankratz (1983)
Description	housing permits data	coal production data
n	84	96
model	(2,0,2)	(2,0,3)
$\hat{\sigma}_a^2$	52.5683	9.599×10^6
$\bar{y}$	108.1274	3.747×10^4

Table 5.3: Table of Results for Example 1

			Profiling		Linearization	
$g(\boldsymbol{\theta})$	MLE	se	L.B.	U.B.	L.B.	U.B.
$\hat{\phi}_1$	-1.3027	0.22	-1.5916	-0.7438	-1.7322	-0.8732
$\hat{\phi}_2$	0.6079	0.17	0.1184	0.8554	0.2602	0.9555
$\hat{\theta}_1$	-0.0939	0.25	-0.4710	0.4950	-0.5919	0.4041
$\hat{\theta}_2$	0.3949	0.15	0.1195	0.6780	0.1040	0.6857

Table 5.4: Table of Results for Example 2

			Profiling		Linearization	
$g(\boldsymbol{\theta})$	MLE	se	L.B.	U.B.	L.B.	U.B.
$\hat{\phi}_1$	-1.8508	0.07	-1.9569	-0.3666	-1.9908	-1.7108
$\hat{\phi}_2$	0.8647	0.08	0.6830	0.9701	0.7247	1.0047
$\hat{\theta}_1$	-1.4397	0.12	-1.6716	0.2693	-1.6725	-1.2069
$\hat{\theta}_2$	0.6379	0.19	0.3085	0.9747	0.3073	0.9685
$\hat{\theta}_3$	-0.1880	0.13	-0.4003	0.0426	-0.4053	0.0293

5.11 Nomenclature

a_t	= white noise sequence
B	= backshift operator
c	= a constant
$cov(\hat{\boldsymbol{\theta}})$	= variance covariance matrix of $\hat{\boldsymbol{\theta}}$
e_t	= sequence of correlated residuals from an estimated model
$E[\cdot]$	= expected value operator
$F(p, n-p; \alpha)$	= upper α quantile for the F distribution with p and $n-p$ degrees of freedom
$g(\boldsymbol{\theta})$	a function of parameters
$\mathcal{I}$	= information matrix
$L(\boldsymbol{\theta})$	= likelihood function evaluated at $\boldsymbol{\theta}$
$\mathcal{L}(\boldsymbol{\theta})$	= natural logarithm of the likelihood function of $\boldsymbol{\theta}$
LR	= likelihood ratio
$\mathcal{LR}$	= natural logarithm of the likelihood ratio
n	= number of observations
p	= number of estimated parameters
s	= estimated standard deviation of the random errors
$S(\boldsymbol{\theta})$	$= \boldsymbol{y_n}'\boldsymbol{\Omega}_n^{-1}\boldsymbol{y_n}$
se	= standard error
$t(n-p; \alpha/2)$	= upper $\alpha/2$ quantile of the t distribution with $n-p$ degrees of freedom
t_{max}	= time at which the concentration of Species A reaches a maximum , (weeks)
$v(u)$	= covariance at lag u

$\boldsymbol{V}$ = $n \times p$ matrix of elements v_{ij} representing the first derivative of $f(\boldsymbol{x_i}, \boldsymbol{\theta})$ with respect to the j^{th} parameter

w_t = a time series

x_t = a time series

$\boldsymbol{x}$ = $1 \times m$ row vector of m independent variables

$\boldsymbol{X}$ = $n \times p$ matrix of elements x_{ij} representing the level of the j^{th} independent variable for observation i

y = response variable

$\boldsymbol{y_n}$ = $n \times 1$ column vector of values of the response variable

z = forward shift operator

z_t = a time series transformed using Ansley's transformation

Greek letters

α = significance level

δ = the delta operator

$\delta(\theta_i)$ = studentized value of θ_i

$\delta(g(\boldsymbol{\theta}))$ = studentized value of $g(\boldsymbol{\theta})$

Δ = time delay

ϵ = additive random error

$\boldsymbol{\epsilon}$ = $n \times 1$ column vector of random errors

θ_i = i^{th} parameter of a model

$\theta(B)$ = moving average polynomial of a time series model

$\boldsymbol{\theta}$ = $p \times 1$ column vector of parameters of a model

$\nu(u)$ = covariance at lag u

$\rho(u)$ = covariance at lag u

σ_a^2 = variance of the white noise sequence a_t

$\tau(\theta_i)$	= profile t statistic for θ_i
$\tau(g(\boldsymbol{\theta}))$	= profile t statistic for $g(\boldsymbol{\theta})$
Υ	= expected value of $\mathcal{I}$
$\phi(B)$	= autoregressive polynomial of a time series model
$\chi^2(1)$	= the chi-squared distribution with 1 degree of freedom
$psi(t, \boldsymbol{\theta})$	= $p \times 1$ vector of derivatives of a_t with respect to $\boldsymbol{\theta}$
$\boldsymbol{\Omega}_{n,z}$	= variance covariance matrix for z_t
$\sigma_a^2 \boldsymbol{\Omega}_n$	= $n \times n$ covariance matrix for $\boldsymbol{y_n}$

Superscripts

$*$	= a true value
$\hat{}$	= a maximum likelihood estimate
$\tilde{}$	= a constrained estimate
d	= a parameter of a delta model

Abbreviations

AR	autoregressive
ARMA	autoregressive moving average
ARIMA	autoregressive integrated moving average
iid	independently and identically distributed
MA	moving average
MLE	maximum likelihood estimate
PACF	partial autocorrelation function

Chapter 6

Two New Empirical Measures of Nonlinearity

6.1 Abstract

Several measures of nonlinearity have been proposed in the literature (Beale, 1960; Bates and Watts, 1980; Cook and Goldberg 1986; Clarke, 1987; Kang and Rawlings, 1998). The drawbacks of these measures include the computational complexity of the expressions, and their potential unreliability (Cook and Witmer, 1985; van Ewijk and Hoekstra, 1994). We propose a quick and easy measure of nonlinearity for autoregressive moving average (ARMA) models which is based on the proximity of the vector of parameters to a stability/invertibility boundary. For nonlinear regression models, we propose a pseudo-profiling algorithm as a means by which to estimate qualitatively the nonlinearity of a function of the parameters of a proposed model. Both of these approaches are computationally simpler than the measures proposed previously, yet they reliably indicate the degree of nonlinearity to be expected in a function of parameters.

6.2 Introduction

For nonlinear models it is common to compute inference results for any function of the parameters in the model using linear approximations to the model and to the function of parameters of interest. This approach is appealing because it is computationally simple; however, inference results based on this method can be misleading (Bates and Watts, 1988; Donaldson and Schnabel, 1987; Ross, 1990; Lam and Watts, 1991). Measures of nonlinearity can indicate when it is appropriate to use linearization inference results. The motivation for these measures is a desire to avoid the computational burden of iterative approaches to computing inference intervals. However, computing power today is such that, in many cases, the time required to compute iterative inference results is insignificant in terms of the overall modeling and analysis effort. Profiling (Bates and Watts, 1988; Chen, 1991; Chen and Jennrich, 1996; Quinn et al., 1999a; Chapter 3), is an iterative approach to computing inference intervals, which provides both reliable likelihood intervals and graphical displays of the nonlinearity of the inference results (Chen and Jennrich, 1996). We favor the use of profiling whenever possible. Nonetheless, there may be cases in which profiling would be an expensive undertaking; one such case would be a time series having thousands of data points. In these cases there is a need for measures of nonlinearity which are readily and easily computed.

The measures of nonlinearity previously proposed in the literature are complicated and require expressions for the second derivatives of both the model and the function of parameters of interest. For very long time series the Bates and Watts measures require the manipulation of large three-dimensional arrays. Also, the Bates and Watts measures require an iterative search procedure for the maximum curvature. The barriers to the use of these measures are therefore relatively high. Thus, a simple yet reliable indicator of nonlinearity is an attractive alternative.

The paper proceeds as follows. First we review some of the relevant theory and

literature. Then we develop a measure of nonlinearity specifically for use in time series modeling. The measure is based on the proximity of the parameter vector to a stability/invertibility boundary. This distance has previously been found to be a leading contributor to the nonlinearity of the parameters of time series models (Lam and Watts, 1991; Quinn et al., 1999c; Chapter 5). Subsequently, we propose a graphical alternative to profiling which avoids the burden of solving a series of optimization problems. Although the exact statistical interpretation of the results of this method is not as readily identified as for profiling, the qualitative information provided is a reliable indication of the nonlinearity to be expected. This method is proposed for use with nonlinear regression models, although it may also be used with other classes of models, including time series models. The paper concludes with a review of the merits and limitations of the newly proposed measures of nonlinearity.

6.3 Some Background on Measures of Nonlinearity

The Bates and Watts (1980) measures of nonlinearity were developed for nonlinear regression models having the form

$$y = f(\boldsymbol{x}, \boldsymbol{\theta}) + \epsilon \tag{6.1}$$

where y is a response variable, $\boldsymbol{x}$ is a vector of regressor variables, $\boldsymbol{\theta}$ is a vector of all unknown parameters, and ϵ is an additive random error term which is assumed to be independently and identically normally distributed with zero mean and constant variance for all observations. These measures of nonlinearity are based on a geometric concept of curvature. Underlying the method is a decomposition of the matrix of second derivatives of the model with respect to the parameters. Bates and Watts dis-

tinguished between two types of curvature: intrinsic curvature and parameter-effects curvature. As the vector of parameters $\boldsymbol{\theta}$ is changed, the model function $f(\boldsymbol{x}, \boldsymbol{\theta})$ traces out an r-dimensional surface called an expectation surface (Bates and Watts, 1980), also called a solution surface, where r is the total number of parameters to be estimated. When $f(\boldsymbol{x}, \boldsymbol{\theta})$ is linear, the solution surface is planar. When $f(\boldsymbol{x}, \boldsymbol{\theta})$ is nonlinear, the solution surface has curvature. Intrinsic nonlinearity is the extent to which the solution locus deviates from a plane tangent to its surface at the maximum likelihood estimate of the vector of parameters, $\hat{\boldsymbol{\theta}}$. This type of curvature is independent of the parameterization of the model. However, the parameterization of the model does affect the mapping of points on the solution surface to points in the parameter space. For linear models, curves of constant θ_i trace out straight parallel lines on the solution plane (i.e., the mapping from the parameter space is uniform). For nonlinear models, the mapping is not uniform and therefore the curves of constant θ_i are not straight, nor parallel, nor equi-spaced. Parameter-effects nonlinearity refers to the nonlinearity of the mapping of coordinates from the solution surface to the space of the parameters.

To quantify each type of nonlinearity, Bates and Watts decomposed the acceleration vector, defined by the Hessian matrix, into three components. To define the acceleration vector, define the following arbitrary straight line in the space of the parameters passing through the point $\boldsymbol{\theta}_0$:

$$\boldsymbol{\theta}(b) = \boldsymbol{\theta}_0 + b\boldsymbol{h} \tag{6.2}$$

where $\boldsymbol{h}$ is any nonzero $r \times 1$ vector (Bates and Watts, 1980). The matrix of the velocity vectors is

$$\{\boldsymbol{V}\}_{ij} = \frac{\partial f(\boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \theta_j} \tag{6.3}$$

and the Hessian is

$$\{\boldsymbol{H}\}_{ijk} = \frac{\partial^2 f(\boldsymbol{x}_i, \boldsymbol{\theta})}{\partial\theta_j \partial\theta_k} \tag{6.4}$$

where $\boldsymbol{H}$ is an $n \times r \times r$ array. Then, a tangent vector to the solution surface at $b = 0$ is:

$$\dot{\eta}_h = \left.\frac{df(\boldsymbol{x}, \boldsymbol{\theta})}{db}\right|_{\boldsymbol{\theta}_0} = \boldsymbol{V}\boldsymbol{h} \tag{6.5}$$

and the acceleration of line (6.2) is:

$$\ddot{\eta}_h = \left.\frac{d^2 f(\boldsymbol{x}, \boldsymbol{\theta})}{db^2}\right|_{\boldsymbol{\theta}_0} = \boldsymbol{h}^T \boldsymbol{H}\boldsymbol{h} \tag{6.6}$$

Since $\boldsymbol{H}$ is an array we follow the conventions for denoting the two possible types of matrix multiplication used by Bates and Watts (1980, 1988). Square brackets indicate that the summation is over the first index of the array. Multiplication of arrays not contained within square brackets is such that if $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{C}$, and $\boldsymbol{B}$ is an $a \times b \times c$ array, then $\boldsymbol{C}$ will multiply each of the $(b \times c)$ faces of the array (see Bates and Watts (1988) for further details). The decomposition of the acceleration vector can be written as:

$$\ddot{\eta}_h = \ddot{\eta}_h^N + \ddot{\eta}_h^P + \ddot{\eta}_h^G \tag{6.7}$$

where $\ddot{\eta}_h^N$ is normal to the tangent plane, $\ddot{\eta}_h^P$ is parallel to $\dot{\eta}_h$, and $\ddot{\eta}_h^G$ is parallel to the tangent plane normal to $\dot{\eta}_h$ (Bates and Watts, 1980). Then the intrinsic nonlinearity is defined to be:

$$\gamma_h^N = \frac{\|\ddot{\eta}_h^N\|}{\|\dot{\eta}_h\|^2} s\sqrt{p} \tag{6.8}$$

and the parameter effects curvature is defined to be:

$$\gamma_h^T = \frac{\|\ddot{\eta}_h^P + \ddot{\eta}_h^G\|}{\|\dot{\eta}_h\|^2} s\sqrt{p} \tag{6.9}$$

To judge the degree of nonlinearity, Bates and Watts recommended that the scaled curvature be compared to $1/\sqrt{F(r, n-r, \alpha)}$. If γ_h^N and γ_h^T are both less than this critical value the nonlinearity is deemed minor or insignificant.

The intrinsic curvature and the parameter-effects curvature are defined in terms of a directed line. Then, Bates and Watts defined two global measures of nonlinearity, namely maximum intrinsic and parameter-effects curvatures, where

$$\gamma_{max}^N = \max_h \gamma_h^N \tag{6.10}$$

and

$$\gamma_{max}^T = \max_h \gamma_h^T \tag{6.11}$$

and the root mean square intrinsic and parameter-effects curvatures, where

$$(\gamma_{RMS}^N)^2 = \frac{\int_{\|h\|} (\gamma_h^N)^2 dS}{\int_{\|h\|} dS} \tag{6.12}$$

and

$$(\gamma_{RMS}^T)^2 = \frac{\int_{\|h\|} (\gamma_h^T)^2 dS}{\int_{\|h\|} dS} \tag{6.13}$$

where dS is an element of the surface area (Ravishanker, 1994). The relationship between these measures and Beale's (1960) measures of nonlinearity and Box's (1971) measure of bias have been established (Bates and Watts, 1980).

These global measures of nonlinearity proposed by Bates and Watts indicate the

maximum and average curvatures of the solution surface in a region about the maximum likelihood estimates (MLE) of the parameters. However, interest in curvature is often associated with particular parameters or functions of parameters. In such cases, the interest is not in the overall curvature, but in the curvature along directions specified by the function of parameters. Cook and Goldberg (1986), Clarke (1987), and Kang and Rawlings (1998) have developed expressions for marginal curvature measures which indicate the nonlinearity associated with a parameter or function of parameters of interest.

The marginal measure developed by Kang and Rawlings (1998) can be shown to be a generalization of the measures of Cook and Goldberg (1986) and Clarke (1987), and so only the most general expression will be given here. Some notation is needed for the development. The matrix $\boldsymbol{V}$ is subjected to a QR decomposition such that:

$$\boldsymbol{V} = \boldsymbol{Q}_1 \boldsymbol{R}_1 \tag{6.14}$$

where $\boldsymbol{Q}_1$ is an $n \times p$ matrix of orthogonal columns, and $\boldsymbol{R}_1$ is a $p \times p$ upper triangular matrix. In terms of the QR decomposition of $\boldsymbol{V}$, the parameter-effects curvature array of Bates and Watts (1980) is:

$$\boldsymbol{A} = [\boldsymbol{Q}_1^T][\boldsymbol{R}_1^{-T} \boldsymbol{H} \boldsymbol{R}_1^{-1}] \tag{6.15}$$

(Kang and Rawlings, 1998). The array of marginal curvatures proposed by Clarke (1987) is defined by:

$$\boldsymbol{\Gamma} = \boldsymbol{R}_1^{-1}[\boldsymbol{R}_1^{-1}][\boldsymbol{A}]\boldsymbol{R}_1^{-T} \tag{6.16}$$

and the marginal curvature is:

$$m_i = -\frac{1}{2}\left(v_{ii}\right)^{-3/2}\gamma_i s^2 \tag{6.17}$$

where v_{ii} is the ii^{th} element of $(\boldsymbol{V}^T\boldsymbol{V})^{-1}$, γ_i is the iii^{th} element of $\boldsymbol{\Gamma}$, and s^2 is the estimated variance of the additive noise.

Kang and Rawlings generalized this result to apply to any function of parameters $\boldsymbol{\phi} = g(\boldsymbol{\theta})$ as follows. Let

$$\boldsymbol{V}^{\phi} = \frac{\partial f(\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\phi}^T} = \boldsymbol{V}^{\theta}\boldsymbol{D}^{\theta} \tag{6.18}$$

and

$$\boldsymbol{H}^{\phi} = \frac{\partial^2 f(\boldsymbol{x},\boldsymbol{\theta})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^T} = \boldsymbol{D}^{\theta}\boldsymbol{H}\boldsymbol{D}^{\theta} + [\boldsymbol{V}^{\theta}][\boldsymbol{B}^{\theta}] \tag{6.19}$$

where $\boldsymbol{D}^{\theta} = \frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\phi}^T}$ and $\boldsymbol{B}^{\theta} = \frac{\partial^2\boldsymbol{\theta}}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^T}$. Then let $\boldsymbol{D}^{\phi} = \frac{\partial\boldsymbol{\phi}}{\partial\boldsymbol{\theta}^T}$, $\boldsymbol{B}^{\phi} = \frac{\partial^2\boldsymbol{\phi}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$,

$$\boldsymbol{G}^{\theta} = \left((\boldsymbol{V}^{\theta})^T\boldsymbol{V}^{\theta}\right)^{-1} \tag{6.20}$$

$$\tag{6.21}$$

and

$$\begin{aligned}\boldsymbol{G}^{\phi} &= \left((\boldsymbol{V}^{\phi})^T\boldsymbol{V}^{\phi}\right)^{-1} \\ &= \boldsymbol{D}^{\phi}\boldsymbol{G}^{\theta}(\boldsymbol{D}^{\phi})^T\end{aligned} \tag{6.22}$$

Then

$$\begin{aligned} \boldsymbol{\Gamma}^{\boldsymbol{\phi}} &= \boldsymbol{\Gamma}_{\mathbf{1}} - \boldsymbol{\Gamma}_{\mathbf{2}} \\ &= \boldsymbol{D}^{\boldsymbol{\phi}}[\boldsymbol{D}^{\boldsymbol{\phi}}][\boldsymbol{\Gamma}^{\boldsymbol{\theta}}](\boldsymbol{D}^{\boldsymbol{\phi}})^{T} - \boldsymbol{D}^{\boldsymbol{\phi}}\boldsymbol{G}^{\boldsymbol{\theta}}\boldsymbol{B}^{\boldsymbol{\phi}}\boldsymbol{G}^{\boldsymbol{\theta}}(\boldsymbol{D}^{\boldsymbol{\phi}})^{T} \end{aligned} \tag{6.23}$$

Therefore, the marginal curvature for ϕ_i is:

$$m_i^{\phi} = -\frac{1}{2}\left(v_{ii}^{\phi}\right)^{-3/2}\gamma_i^{\phi}s^2 \tag{6.24}$$

where $\gamma_i^{\phi} = \boldsymbol{\Gamma}^{\phi}{}_{iii}$ (Kang and Rawlings, 1998).

Cook and Witmer (1985) and van Ewijk and Hoekstra (1994) did not find a strong relationship between the marginal curvatures for the individual parameters of the logistic regression model and the reliability of the linear approximation confidence intervals for those parameters. Given the sometimes poor performance of measures of nonlinearity, together with the complexity of their expressions and the computational time required for their evaluation, there is a need for a "quick and easy" measure of nonlinearity.

Another graphical means of displaying the nonlinearity associated with a function of parameters $g(\boldsymbol{\theta})$ is the profile t plot (Bates and Watts, 1988; Chen, 1991; Lam and Watts, 1991; Severini and Staniswalis, 1994; Chen and Jennrich, 1996, Quinn et al., 1999a; Chapter 3; Quinn et al., 1999b; Chapter 4). Chen (1991) developed a method for obtaining profiling results for an arbitrary function $g(\boldsymbol{\theta})$ by posing the problem in terms of a constrained optimization:

Maximize

$$L(\boldsymbol{\theta}) \tag{6.25}$$

subject to the constraint

$$g(\boldsymbol{\theta}) = c$$

where $L(\boldsymbol{\theta})$ is the likelihood function, and c is a constant. To profile $g(\boldsymbol{\theta})$ is to solve a series of these optimization problems for a range of values of c greater than and less then the maximum likelihood estimate of $g(\boldsymbol{\theta})$. A profile t plot is then a plot of

$$\tau(g(\tilde{\boldsymbol{\theta}})) = sign(g(\tilde{\boldsymbol{\theta}}) - g(\hat{\boldsymbol{\theta}}))\sqrt{-2\ln\frac{L(\tilde{\boldsymbol{\theta}})}{L(\hat{\boldsymbol{\theta}})}} \tag{6.26}$$

versus $g(\tilde{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$, and $\tilde{\boldsymbol{\theta}}$ is the location of the solution of the optimization problem (6.25) for a specific value of c. Profiling an individual parameter is simply a special case where $g(\boldsymbol{\theta}) = \theta_i$. Profile t plots are also sometimes constructed as plots of $\tau(g(\tilde{\boldsymbol{\theta}}))$ versus $\delta(g(\tilde{\boldsymbol{\theta}}))$, where

$$\delta(g(\tilde{\boldsymbol{\theta}})) = \frac{g(\tilde{\boldsymbol{\theta}}) - g(\hat{\boldsymbol{\theta}})}{se(g(\hat{\boldsymbol{\theta}}))} \tag{6.27}$$

$se(g(\hat{\boldsymbol{\theta}}))$ is the standard error of $g(\hat{\boldsymbol{\theta}})$, and is computed using the expression

$$se(\hat{g}) = s\sqrt{\frac{dg}{d\boldsymbol{\theta}}^T\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}(\boldsymbol{V}^T\boldsymbol{V})^{-1}\frac{dg}{d\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}} \tag{6.28}$$

where $\frac{dg}{d\boldsymbol{\theta}}$ is the $r\times 1$ vector of partial derivatives of $g(\boldsymbol{\theta})$ with respect to the parameters, $\boldsymbol{V}^T\boldsymbol{V}$ is the observed Fisher information matrix, $s = \sqrt{\frac{S(\hat{\boldsymbol{\theta}})}{n-r}}$ is an estimate of the standard deviation of the random errors, and

$$\begin{aligned} S(\boldsymbol{\theta}) &= \sum_{i=1}^{n}(y_i - \boldsymbol{f}(\boldsymbol{x}_i, \boldsymbol{\theta}))^2 \\ &= (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}))^T(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta})) \end{aligned} \tag{6.29}$$

$$= \boldsymbol{e}^T\boldsymbol{e}$$

where $\boldsymbol{e}^T = (e_1, e_2, \ldots, e_n)$ is the vector of residuals.

When the model and the function $g(\boldsymbol{\theta})$ are linear in the parameters, the profile t plot for $g(\boldsymbol{\theta})$ will be a straight line. When either the model or the function of parameters, or both, are nonlinear, then the profile t plot will be curved. The departure of the profile t plot from a straight line is an indicator of nonlinearity (Chen and Jennrich, 1996).

The profile t plot is an attractive way to judge nonlinearity qualitatively. However, because it involves solving a series of constrained optimization problems, it can be computationally intensive. Whereas likelihood intervals are based on (6.26), linearization intervals are based on (6.27). The limits of the linearization confidence interval for $g(\boldsymbol{\theta})$ are

$$g(\hat{\boldsymbol{\theta}}) \pm st(n-p;\alpha/2)\sqrt{\frac{\partial g}{\partial \boldsymbol{\theta}}^T(\boldsymbol{V}^T\boldsymbol{V})^{-1}\frac{\partial g}{\partial \boldsymbol{\theta}}} \tag{6.30}$$

where $t(n-p, \alpha/2)$ is the upper $\alpha/2$ quantile of the t distribution with $n-p$ degrees of freedom. The linearization confidence interval is a popular means of expressing uncertainty because the analytic expression for the limits makes the computational effort minimal. However, these inference intervals can be unreliable and even misleading when the model and/or the function of parameters is highly nonlinear. Measures of nonlinearity can be used to decide whether linearization inference results will be adequate.

6.4 A Measure of Nonlinearity for ARMA models

Here, we propose an alternate measure of nonlinearity specifically for time series models. Our measure is based on the fact that the nonlinearity of the parameters of time series models has been found to be closely related to the proximity of the parameter vector to the nearest stability/invertibility boundary (Lam and Watts, 1991). The new measure also has the advantage that its expected value is easily obtained in the absence of data.

Consider an ARMA(p,q) model of the form:

$$\phi(B)y_t = \theta(B)a_t \tag{6.31}$$

where $\phi(B) = (1+\phi_1 B+\ldots+\phi_p B^p)$, $\theta(B) = (1+\theta_1 B+\ldots+\theta_q B^q)$, B is the backshift operator defined as $By_t = y_{t-1}$, $\{y_t\}$ is an observed stationary stochastic process with mean zero, and $\{a_t\}$ is a normally distributed white noise process such that all a_t are independently and identically normally distributed (i.e., $\{a_t\}$ is a sequence of iid $N(0,\sigma_a^2)$ random variables). For a time series model, $\boldsymbol{\theta}^T = (\phi_1,\ldots,\phi_p,\theta_1,\ldots,\theta_q)$. The total number of parameters to be estimated is $r = p + q$, where p and q are the number of parameters in the autoregressive (AR) and moving average (MA) polynomials, respectively. Although σ_a^2 is usually unknown, we estimate it based on the modified residuals of the fitted model (see Lam and Watts, 1991) and we do not include it in $\boldsymbol{\theta}$. Note that a nonstationary or non-zero mean time series may be transformed to conform to the above model by first appropriately differencing or mean centering the data, respectively.

Figure 6.1 illustrates the stability/invertibility region for the parameters of an AR(2) or an MA(2) model. These stability limits are well defined and easily shown

for the 2-dimensional case (Box and Jenkins, 1976; Wei, 1990); however, in higher dimensions, the stability/invertibility boundaries are not so readily identifiable nor are they easily illustrated. To define a measure of nonlinearity, we first transform the

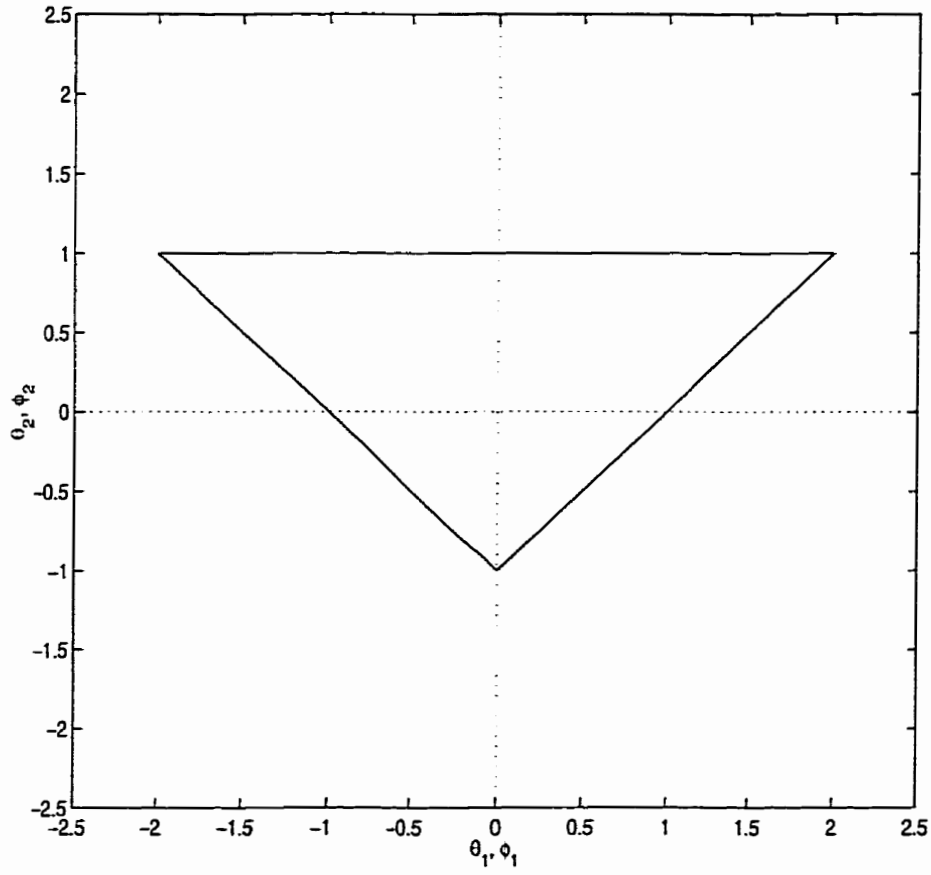

Figure 6.1: The stability/invertibility region for the parameters of an AR(2) or MA(2) polynomial is the interior of the triangle.

parameters of the AR polynomial and then the parameters of the MA polynomial using the definition of the partial autocorrelation function (PACF) for AR processes.

The PACF function is defined by

$$\phi_{kk} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_2 \\ \vdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & \rho_k \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_{k-2} \\ \vdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & 1 \end{vmatrix}} \tag{6.32}$$

where ϕ_{kk} is the k^{th} partial autocorrelation and ρ_k is the k^{th} autocorrelation (Wei, 1990). PACF values may be computed efficiently using the Levinson algorithm (Abraham and Ledolter, 1983). We apply the definition of the partial autocorrelation function for an AR process separately to the AR and MA components of the ARMA model such that the resulting k^{th} parameter of the MA or AR polynomial will be equal to ϕ_{kk} (as defined for AR processes). That is, we treat each polynomial as if it were a polynomial defining an AR process, and then find the values of the PACF on this basis. The parameter vector $\boldsymbol{\phi}$ consists of all r PACF values computed in this way. The advantage of this transformation is that the individual stability/invertibility limits for the new parameters are ± 1, and the r-dimensional stability region is a hypercube with sides located at ± 1 (Åström and Wittenmark, 1990). This makes it easy to compute the distance of a vector of parameters $\boldsymbol{\phi}$ from the closest stability/invertibility boundary. Also, 2-dimensional projections of the stability region can be easily used to visualize the position of the r-dimensional vector $\boldsymbol{\phi}$.

Define $\boldsymbol{\phi}_{95,i,1}$ to be the vector of conditional maximum likelihood estimates of all of the parameters (in PACF space) when the parameter θ_i is at the upper limit of its 95 % linearization confidence interval, let $\boldsymbol{\phi}_{95,i,2}$ be the vector of values of all of the parameters (in PACF space) when the parameter θ_i is at the lower limit of its 95 % linearization confidence interval, and let $\hat{\boldsymbol{\phi}}$ represent the vector of unconditional maximum likelihood estimates of the parameters in PACF space. Let $\boldsymbol{\phi}_{95,i,j}$ represent either $\boldsymbol{\phi}_{95,i,1}$ or $\boldsymbol{\phi}_{95,i,2}$. The expressions for the values of the parameters at the limits of the linearization confidence interval for θ_i are given later in (6.38) and (6.39). Assume that the vector joining $\boldsymbol{\phi}_{95,i,j}$ and $\hat{\boldsymbol{\phi}}$ provides a good indication of the direction in which the parameter vector moves as the parameter θ_i is profiled. Experience has shown that this assumption is excellent when $\hat{\boldsymbol{\phi}}$ and $\boldsymbol{\phi}_{95,i,j}$ are "far" from the stability/invertibility boundaries. The assumption breaks down when $\boldsymbol{\phi}_{95,i,j}$ is outside (or just within) the stability/invertibility region. We propose that the distance from $\hat{\boldsymbol{\phi}}$ to the nearest stability boundary in the direction $\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_{95,i,j}$, scaled by the distance from $\hat{\boldsymbol{\phi}}$ to $\boldsymbol{\phi}_{95,i,j}$, is a useful measure of nonlinearity. An expression for this measure is developed below.

First, identify the conditional maximum likelihood estimates of the $r-1$ parameters when one parameter θ_i is at one of the limits of its confidence interval as determined using the linearization approach. The boundary of the joint $(1-\alpha)100$ % linearization confidence region for the parameters is defined by:

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \boldsymbol{V}^T \boldsymbol{V} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = r s^2 F(1, n-r; \alpha) \tag{6.33}$$

where $F(1, n-r; \alpha)$ is the upper α quantile of the F distribution with 1 and $n-r$ degrees of freedom. This expression defines an ellipse whose principal axes are the eigenvectors of $\boldsymbol{V}^T\boldsymbol{V}$. The locations of the limits of the linearization confidence interval for each parameter lie along one of these eigenvectors. Note that when individual

confidence intervals, as opposed to joint confidence regions, are of interest, the appropriate scaling of the contours is $s^2F(1, n-r; \alpha)$ and not $rs^2F(r, n-r; \alpha)$ (Donaldson and Schnabel, 1987) . Also, the limit of a $(1-\alpha)100$ % confidence interval for an individual parameter θ_i can be obtained from

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = \lambda(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{a} \tag{6.34}$$

where $\boldsymbol{a} = \frac{\partial\theta_i}{\partial\boldsymbol{\theta}} = [0, \ldots, 0, 1, 0, \ldots, 0]^T$ and the 1 is in row i, and λ is a constant which defines the confidence level. Expression (6.34) can also be used to locate the limits of a linearization confidence interval for a function of the parameters, $g(\boldsymbol{\theta})$, by letting $\boldsymbol{a} = \frac{\partial g}{\partial\boldsymbol{\theta}}$. Substituting (6.34) into (6.33):

$$\begin{aligned} \lambda\boldsymbol{a}^T(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{V}^T\boldsymbol{V}\lambda(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{a} &= s^2F(1, n-r; \alpha) \qquad (6.35) \\ \lambda^2\boldsymbol{a}^T(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{a} &= s^2F(1, n-r; \alpha) \end{aligned}$$

Therefore,

$$\lambda = \sqrt{\frac{s^2F(1, n-r; \alpha)}{\boldsymbol{a}^T(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{a}}} \tag{6.36}$$

$$= \frac{s^2\sqrt{F(1, n-r; \alpha)}}{se(g(\hat{\boldsymbol{\theta}}))} \tag{6.37}$$

where $se(g(\hat{\boldsymbol{\theta}})) = s\sqrt{\boldsymbol{a}^T(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{a}}$ is the standard error of $g(\hat{\boldsymbol{\theta}})$.

It has been our experience that the profile t plot for an individual parameter will be significantly nonlinear if, during the profiling process, the vector of parameters approaches a stability/invertibility boundary. This is consistent with the work of Lam and Watts (1991). As a parameter θ_i is profiled, the values of the remaining $(r-1)$ parameters adjust so as to compensate for the change in the value of θ_i. If, upon this adjustment, the vector of the $(r-1)$ remaining parameters approaches its

stability/invertibility boundary, then the profile t plot for θ_i will become nonlinear.

We illustrate the concepts discussed above by considering a constructed example in which an MA(2) model has parameter estimates $\hat{\boldsymbol{\theta}}^T = (-1.2,\ 0.3)$. First, reparameterizing the model using the PACF transformation defined in (6.32) yields the corresponding estimates $\hat{\boldsymbol{\phi}}^T = (0.9231,\ -0.3)$. Then, the upper limit of the 95 % linearization confidence interval for θ_1 is:

$$\boldsymbol{\theta}_{95,1,1} = \hat{\boldsymbol{\theta}} + \frac{s\, t(n-r, \alpha/2)(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{a}}{\sqrt{\boldsymbol{a}^T(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{a}}} \tag{6.38}$$

where $\boldsymbol{a}^T = (1, 0)$, and the location of the lower limit is:

$$\boldsymbol{\theta}_{95,1,2} = \hat{\boldsymbol{\theta}} - \frac{s\, t(n-r, \alpha/2)(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{a}}{\sqrt{\boldsymbol{a}^T(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{a}}} \tag{6.39}$$

Note that (6.38) and (6.39) provide the conditional maximum likelihood values of the parameters when θ_1 is at the limits of its linearization confidence interval. Then, let $\boldsymbol{\phi}_{95,1,1}$ and $\boldsymbol{\phi}_{95,1,2}$ represent the corresponding limits in PACF space. The limits of the 95 % linearization confidence intervals for the parameters of this constructed example are given in Table 6.1. Figure 6.2 shows that two of the linear approximation limits lie outside of the stability/invertibility boundary. This would not be the case for the corresponding likelihood limits since these limits remain true to the stability/invertibility boundaries because the likelihood function decreases sharply as the parameter vector approaches a stability boundary, and a penalty is imposed if the parameters of the MA function move outside of the invertibility region.

In general, it is not simply the distance of the parameter being profiled from its individual stability/invertibility boundary which is ultimately important in determining the nonlinearity of that parameter, but rather, it is the directed distance of $\hat{\boldsymbol{\phi}}$ from a stability boundary along the vector joining $\hat{\boldsymbol{\phi}}$ and $\boldsymbol{\phi}_{95,i,j}$. Let $\boldsymbol{\phi}_{b,i,j}$ be the location

where the the vector from $\hat{\boldsymbol{\phi}}$, passing through $\boldsymbol{\phi}_{95,i,j}$, crosses a stability/invertibility boundary. Referring to Figure 6.2, the distances from $\hat{\boldsymbol{\phi}}$ to each of the $\boldsymbol{\phi}_{b,i,j}$'s may be used as a measure of nonlinearity. The $\boldsymbol{\phi}_{b,i,j}$ are found as follows.

Table 6.1: The Locations of the Expected Confidence Limits for Illustration 1.

$\hat{\boldsymbol{\theta}}$	(-1.2 0.3)
$\hat{\boldsymbol{\phi}}$	(0.9231 -0.3)
$\boldsymbol{\theta}_{95,1,1}$	(-0.1632 -0.6571)
$\boldsymbol{\theta}_{95,1,2}$	(-2.2368 1.2571)
$\boldsymbol{\phi}_{95,1,1}$	(0.4758 0.6571)
$\boldsymbol{\phi}_{95,1,2}$	(0.9910 -1.2571)
$\boldsymbol{\phi}_{b,1,1}$	(0.3156 1.0)
$\boldsymbol{\phi}_{b,1,2}$	(0.9728 -1.0)
$\boldsymbol{\theta}_{95,2,1}$	(-2.1571 1.3368)
$\boldsymbol{\theta}_{95,2,2}$	(-0.2429 -0.7368)
$\boldsymbol{\phi}_{95,2,1}$	(0.9231 -1.3368)
$\boldsymbol{\phi}_{95,2,2}$	(0.9231 0.7368)
$\boldsymbol{\phi}_{b,2,1}$	(0.9231 -1.0)
$\boldsymbol{\phi}_{b,2,2}$	(0.9231 1.0)

Let

$$\xi_{i,j} = \frac{\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_{b,i,j}}{\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_{95,i,j}} \tag{6.40}$$

Since $\boldsymbol{\phi}_{b,i,j}$ will have an element at ± 1, we look for the parameter (in PACF space) which first crosses a stability/invertibility boundary in the direction of the vector $\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_{b,i,j}$. Let

$$\ell = \min \left\{ abs \left(\frac{\hat{\boldsymbol{\phi}} - \boldsymbol{k}}{\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_{95,i,j}} \right) \right\} \tag{6.41}$$

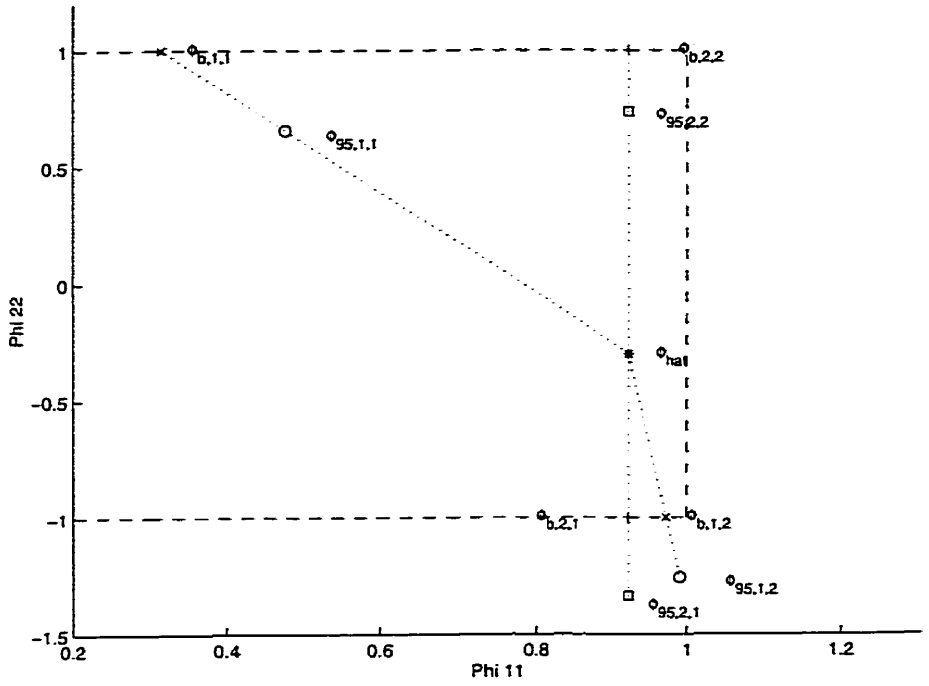

Figure 6.2: An Illustration of the use of the PACF space to estimate nonlinearity. Key: $*$ represents $\hat{\boldsymbol{\phi}}$, $\circ$ represents $\boldsymbol{\phi}_{95,1,i}$, $\square$ represents $\boldsymbol{\phi}_{95,2,i}$, x represents $\boldsymbol{\phi}_{b,1,i}$, and $+$ represents $\boldsymbol{\phi}_{b,2,i}$.

where $\boldsymbol{k}$ is a column of ones and minus ones such that

$$\boldsymbol{k} = sign(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_{b,i,j}) \tag{6.42}$$

Then,

$$\boldsymbol{\phi}_{b,i,j} = \hat{\boldsymbol{\phi}} - \ell(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_{95,i,j}) \tag{6.43}$$

The distance to the boundary is $\|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_{b,i,j}\|$. We propose that the distances from $\hat{\boldsymbol{\phi}}$ to each of the $\boldsymbol{\phi}_{b,i,j}$ can be used as indicators of the anticipated nonlinearity. The measure of nonlinearity $\zeta_{i,j}$ is defined by:

$$\zeta_{i,j} = \frac{\|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_{b,i,j}\|}{\|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_{95,i,j}\|} \tag{6.44}$$

There are two values of ζ for each parameter in the model, one for each of the

two limits of the likelihood interval for that parameter. We define ζ_{min} to be the minimum value of the $2r$ values of ζ. When $\zeta_{min} > 1$, all of the $\phi_{b,i,j}$ and $\phi_{95,i,j}$ are within the stability limits. When $\zeta_{min} < 1$, at least one $\phi_{95,i,j}$ falls outside the stability/invertibility region. The closer ζ_{min} is to zero, the more nonlinearity we anticipate.

Note that the concepts have been illustrated using a 2-dimensional case for ease of display. However, the results readily generalize to higher dimensions since the PACF parameterization is such that the stability region in PACF space will always be an r-dimensional hypercube centered at zero with sides at ± 1.

The expected value of ζ_{min} may be computed in the absence of data. The expected nonlinearity, $\zeta_{min,ap}$, is computed based on (6.44), the only difference being that the locations of the limits of the individual linearization confidence intervals are estimated based on the Cramer-Rao expression for the variance-covariance matrix for the parameter estimates. The expression is developed as follows.

$$cov(\hat{\boldsymbol{\theta}}) \geq \sigma_a^2 [E\{\boldsymbol{\mathcal{I}}\}]^{-1} \tag{6.45}$$

$$\geq \sigma_a^2 \left[\sum_{t=1}^{n} E\left\{ \psi(t,\theta_0) \psi^T(t,\theta_0) \right\} \right]^{-1} \tag{6.46}$$

where $\mathcal{I}$ is the information matrix with elements:

$$\mathcal{I}_{ij} = \frac{\partial \ln L(\boldsymbol{\theta}, \sigma^2)}{\partial \theta_i \partial \theta_j} \tag{6.47}$$

and

$$\psi(t,\theta_0) = \frac{\partial a_t}{\partial \boldsymbol{\theta}} \tag{6.48}$$

For an ARMA(p,q) model, the expressions for the derivatives of a_t with respect to

the parameters are (Ravishanker, 1994):

$$\frac{\partial a_t}{\partial \phi_k} = -\frac{1}{\phi(B)} a_{t-k} = -\frac{1}{\theta(B)} y_{t-k} \tag{6.49}$$

$$\frac{\partial a_t}{\partial \theta_l} = \frac{1}{\theta(B)} a_{t-l} = \frac{\phi(B)}{\theta^2(B)} y_{t-l} \tag{6.50}$$

Let

$$v(u) = cov\left(\frac{\partial a_t}{\partial \phi_k}, \frac{\partial a_{t+u}}{\partial \phi_k}\right) = cov\left(\frac{\partial a_t}{\partial \phi_k}, \frac{\partial a_t}{\partial \phi_{k+u}}\right) \tag{6.51}$$

$$\nu(u) = cov\left(\frac{\partial a_t}{\partial \theta_l}, \frac{\partial a_{t+u}}{\partial \theta_l}\right) = cov\left(\frac{\partial a_t}{\partial \theta_l}, \frac{\partial a_t}{\partial \theta_{l+u}}\right) \tag{6.52}$$

$$\rho_{\phi\theta}(u) = cov\left(\frac{\partial a_t}{\partial \phi_l}, \frac{\partial a_{t+u}}{\partial \theta_l}\right) = cov\left(\frac{\partial a_t}{\partial \phi_l}, \frac{\partial a_t}{\partial \theta_{l+u}}\right) \tag{6.53}$$

(Åström, 1980) and

$$\varrho_{\phi\theta}(u) = cov\left(\frac{\partial a_t}{\partial \phi_k}, \frac{\partial a_{t+u}}{\partial \theta_l}\right) \tag{6.54}$$

Note that

$$\varrho_{\phi\theta}(u) \neq \varrho_{\theta\phi}(u) \tag{6.55}$$

but

$$\varrho_{\phi\theta}(u) = \varrho_{\theta\phi}(-u) \tag{6.56}$$

Then, in order to calculate $Cov(\hat{\boldsymbol{\theta}})$ based on (6.46) so as to obtain a value for $se(g(\hat{\boldsymbol{\theta}}))$,

we develop the expression:

$$\begin{aligned}\Upsilon &= E\{\boldsymbol{\mathcal{I}}\} \\ &= E\{\boldsymbol{\psi}\boldsymbol{\psi}^{\boldsymbol{T}}\} \\ &= \begin{bmatrix} v(0) & \cdots & v(p-1) & \varrho_{\phi\theta}(0) & \cdots & \varrho_{\phi\theta}(q-1) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ v(p-1) & \cdots & v(0) & \varrho_{\phi\theta}(1-p) & \cdots & \varrho_{\phi\theta}(0) \\ \varrho_{\phi\theta}(0) & \cdots & \varrho_{\phi\theta}(1-p) & \nu(0) & \cdots & \nu(q-1) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \varrho_{\phi\theta}(q-1) & \cdots & \varrho_{\phi\theta}(0) & \nu(q-1) & \cdots & \nu(0) \end{bmatrix}\end{aligned} \tag{6.57}$$

Finally, $E\left\{se(g(\hat{\boldsymbol{\theta}}))\right\} = \sigma_a \sqrt{\frac{\partial g}{\partial \boldsymbol{\theta}}^T \Upsilon^{-1} \frac{\partial g}{\partial \boldsymbol{\theta}}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$.

6.4.1 Interpreting ζ_{min}

For an AR process, the derivatives of the model (in terms of the noise a_t) with respect to the parameters are:

$$\frac{\partial a_t}{\partial \phi_i} = -E\{y_{t-i}\} + \boldsymbol{\phi}(B)\frac{\partial E\{y_t\}}{\partial \phi_i} \tag{6.58}$$

When measured values of y_t are available, then $E\{y_t\} = y_t$ and $\frac{\partial E\{y_t\}}{\partial \phi_i} = 0$; therefore, except for the unknown initial conditions, the AR model is linear in the parameters (Box and Jenkins, 1976). Then, for AR models, we expect the starting values to have a negligible impact on $\frac{\partial a_t}{\partial \phi_i}$ when the time series is moderately long. The linearization confidence intervals should be almost exact, except when the vector of parameter

values approaches a stability boundary. For a pure moving average process:

$$\frac{\partial a_t}{\partial \theta_i} = \boldsymbol{\theta}^{-2}(B)E\{y_{t-i}\} + \boldsymbol{\theta}^{-1}(B)\frac{\partial E\{y_t\}}{\partial \theta_i} \tag{6.59}$$

and the model is always nonlinear in the parameters (Box and Jenkins, 1976). For MA models, we expect to observe nonlinearity in most cases except when n is large enough that asymptotic results are appropriate. However, in our experience, the nonlinearity displayed by MA parameters which are not close to an invertibility boundary tend to be moderate. It is only when the parameter values approach an invertibility limit that significant nonlinearity is observed. For mixed models (i.e., ARMA(p,q) models) the derivatives of a_t with respect to the parameters are given by (6.49) and (6.50); therefore, like the MA(q) model, the ARMA(p,q) model is always nonlinear in the parameters except when $q = 0$. It is of interest to know how the presence of AR parameters affects the nonlinearity of the MA parameters and vice versa. The above observations and generalizations indicate a need for a different set of rules by which to judge the nonlinearity indicated by ζ_{min} and $\zeta_{min,ap}$, depending on whether the model is AR, MA or ARMA.

6.5 Illustrative Examples

Ravishanker (1994) computed the Bates and Watts measures of nonlinearity for 16 time series. Data sets 1 to 6 considered by Ravishanker are used here as Examples 1 to 6, respectively. These represent the data sets which were available to us, and which did not involve seasonal effects. We compare the measure of nonlinearity ζ_{min} to the measures of nonlinearity computed by Ravishanker. The results are given in Table 6.2, where it can be seen that the indications of nonlinearity provided by ζ_{min} do not agree with those provided by Ravishanker's the indicators. For example, the relative curvatures γ^T_{max} and γ^N_{max}, and the root mean square curvatures γ^T_{RMS}

and γ^N_{RMS}, which are considered to indicate significant nonlinearity only when their values are greater than 0.3 (Ravishanker, 1994), suggest that Example 6 has relatively low nonlinearity. Note that the measures of Bates and Watts indicate significant nonlinearity when the values of curvature are large (> 0.3), whereas the measure ζ_{min} indicates significant nonlinearity when its values are small (< 1). Since in Example 6 $\zeta_{min} = 0.2984$ is significantly less than 1, severe nonlinearity is indicated. Indeed, the profile t plots for this example show severe nonlinearity (see Figure 6.3). Others, including Cook and Witmer (1984) and van Ewijk and Hoekstra (1994), have also found that the Bates and Watts (1980) measures sometimes fail to show nonlinearity.

Table 6.2: Measures of Nonlinearity for Published Data Sets

Example	n	Model	γ^N_{max}	γ^T_{max}	γ^N_{RMS}	γ^T_{RMS}	ζ_{min}	$\zeta_{min,ap}$
1	197	(1,0,1)	0.4491	6.8639	0.2009	3.0325	1.0435	1.0738
2	396	(0,1,1)	0.1001	0.0261	0.1001	0.0261	8.9430	8.9592
3	226	(0,2,2)	0.2242	0.1968	0.1858	0.1279	6.0986	6.1276
4	90	(0,1,1)	0.2763	0.2662	0.2763	0.2662	1.8388	1.9279
5	84	(2,0,2)	0.6477	0.3646	0.3182	0.1961	1.4107	1.4782
6	96	(2,0,3)	0.1343	0.0542	0.0729	0.0275	0.2984	0.2496

The most significant nonlinearity identified by Ravishanker was for Example 1. The value of ζ_{min} for this example is 1.0435. Because this value of ζ_{min} is only marginally greater than 1, it therefore warns of only moderate nonlinearity. The profile t plots for this example (see Figure 6.4) also show only minor nonlinearities.

The advantage of using ζ_{min} as a measure of nonlinearity is that, in addition to using information about the location of the least squares estimates of the parameters, it also uses information about the path that the parameter vector is likely to follow as the value of an individual parameter or function of parameters is changed. This

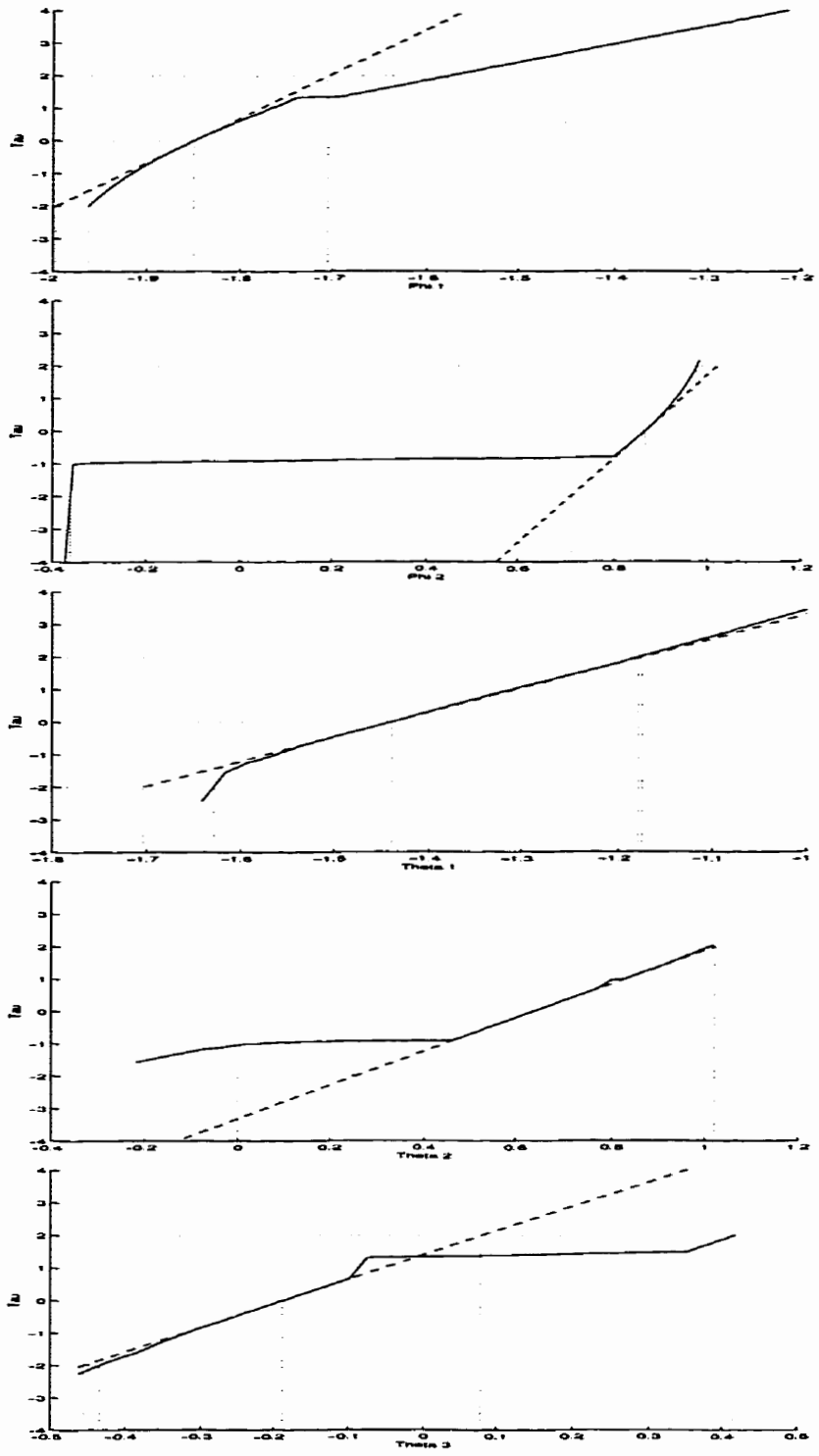

Figure 6.3: Profile t plots for the parameters of Example 6. Reference line: dashed line; Profile t plot: solid line; Dotted lines indicate the maximum likelihood estimates and the limits of the 95 % linearization and likelihood intervals.

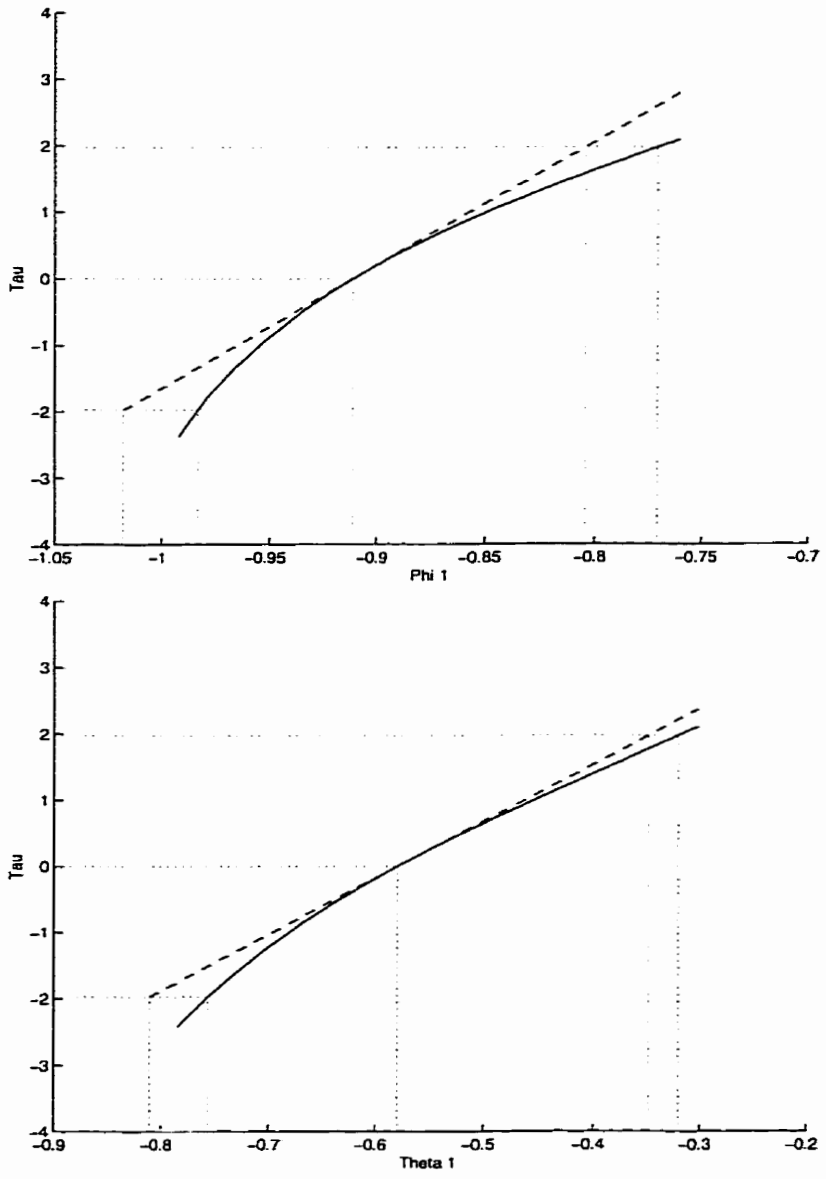

Figure 6.4: Profile t plots for the parameters of Example 1. Reference line: dashed line; Profile t plot: solid line; Dotted lines indicate the maximum likelihood estimates and the limits of the 95 % linearization and likelihood intervals.

information is important in determining the nonlinearity observed in the parameters of time series models because nonlinearity in these cases is intimately related to the proximity of the parameter vector to a stability/invertibility boundary. We emphasize that ζ_{min} is a local measure of nonlinearity in that the scaling of the measure, and the directions along which distances are measured, are based on a linear approximation of the expectation surface centered at $\hat{\boldsymbol{\theta}}$. Likewise, the measures of Bates and Watts (1980) are also local measures in that they use quadratic information about the expectation surface centered at $\hat{\boldsymbol{\theta}}$, and the curvature of the linearization confidence region in the space of the parameters is often used to interpret the results.

Table 6.2 also provides values for the expected nonlinearity, $\zeta_{min,ap}$, for the six examples. The expected nonlinearity predicts the observed nonlinearity well. Plots of the locations of the linearization confidence interval limits for individual parameters in examples 1 and 6 are shown in Figures 6.5 and 6.6, respectively. Also, the locations where the vectors $\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_{95,i,j}$ cross the nearest boundary are indicated. Some of the $\boldsymbol{\phi}_{95,i,j}$ values fall well outside the stability/invertibility region for Example 6, whereas all $\boldsymbol{\phi}_{95}$ values for Example 1 are within the stability/invertibility region, although two are very close to the boundary. These graphical displays are simply another way of presenting the information delivered by ζ_{min}. The plots help to reinforce the geometrical concepts that form the basis for this measure of nonlinearity.

Note that some of the discrepancy between the indications of nonlinearity provided by ζ_{min} and the measures used by Ravishanker for Example 6 may be due to the fact that the likelihood function for this data set has at least two local maxima. If the MLEs of the parameters obtained by Ravishanker were based on an alternate local maximum, then there could be marked differences in the estimates of nonlinearity. The presence of multiple local optima contributes to the high degree of nonlinearity observed in this Example. The maximum likelihood estimates of the parameters $\hat{\boldsymbol{\theta}}$ and the constrained likelihood estimates $\tilde{\boldsymbol{\theta}}$ reported in this paper were computed

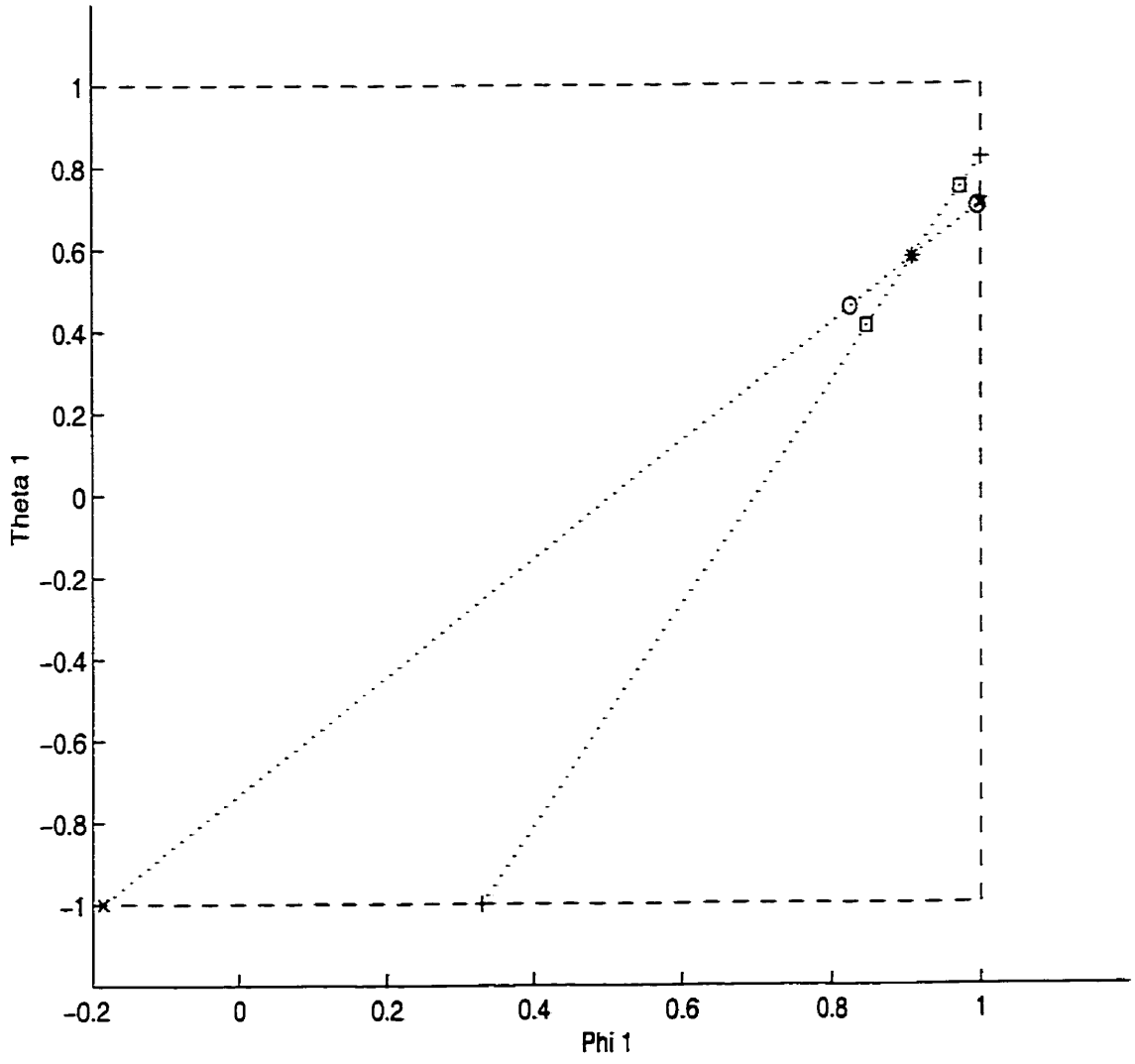

Figure 6.5: Two-dimensional projection plots of the points used to compute ζ_{min} in the PACF space for Example 1. Key: $*$ represents $\tilde{\boldsymbol{\phi}}$, $\circ$ represent $\boldsymbol{\phi}_{95,1}$, $\square$ represent $\boldsymbol{\phi}_{95,2}$, x represent $\boldsymbol{\phi}_{b,1}$, and $+$ represent $\boldsymbol{\phi}_{b,2}$.

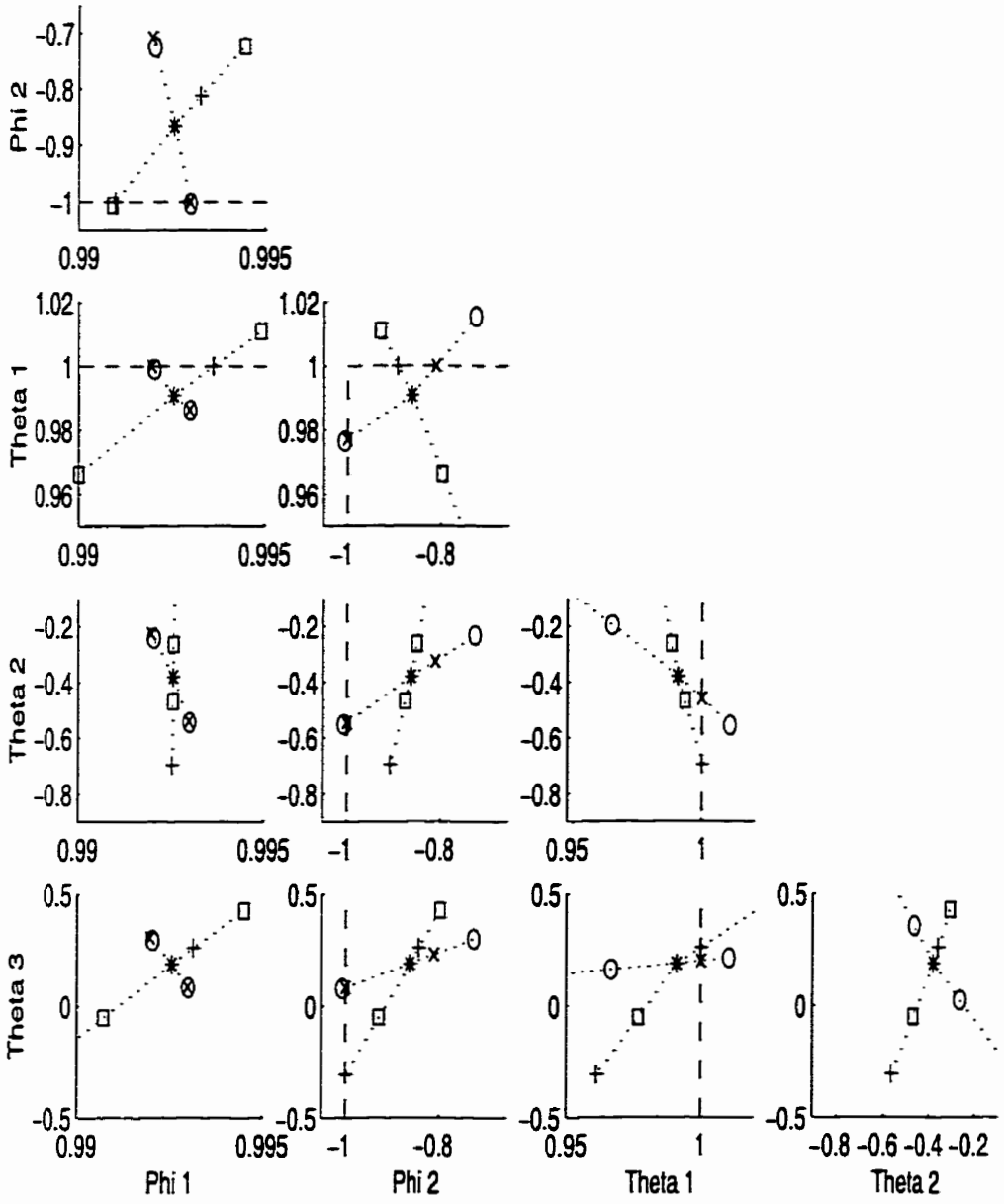

Figure 6.6: Two-dimensional projection plots of the points used to compute ζ_{min} in the PACF space for Example 6. Key: $*$ represents $\tilde{\phi}$, $\circ$ represent $\phi_{95,1}$, $\square$ represent $\phi_{95,2}$, x represent $\phi_{b,1}$, and $+$ represent $\phi_{b,2}$.

using an exact likelihood algorithm proposed by Ansley (1979).

6.6 An alternate measure of nonlinearity for regression models

Profile t plots provide a qualitative measure of nonlinearity for functions of parameters (Chen and Jennrich, 1996). For profile t plots of τ versus $g(\boldsymbol{\theta})$, the deviation of the profile from a straight line having slope $1/se(g(\hat{\boldsymbol{\theta}}))$ and passing through the point $(g(\hat{\boldsymbol{\theta}}), 0)$ is indicative of the nonlinearity of the solution surface in the direction associated with changes in $g(\boldsymbol{\theta})$. This deviation incorporates both intrinsic and parameter-effects nonlinearity.

However, as noted previously, profiling may be computationally intensive because it requires the solution of an optimization problem before the value of τ can be computed for a specified value of $g(\boldsymbol{\theta})$. That is, at each iteration of the profiling algorithm, a search is done for the values $\tilde{\boldsymbol{\theta}}$ which maximize $L(\boldsymbol{\theta})$ given a constraint on $g(\boldsymbol{\theta})$. Then, the profile t plot for $g(\boldsymbol{\theta})$ indicates the nonlinearity of the solution surface in the direction along the locus traced by $\tilde{\boldsymbol{\theta}}$. For linear models with additive independently and identically normally distributed (iid $N(0, \sigma^2)$) noise, the contours of $L(\boldsymbol{\theta})$ are a series of concentric ellipses, and the path traced by $\tilde{\boldsymbol{\theta}}$ will be a straight line if $g(\boldsymbol{\theta})$ is also linear. For the case of an individual parameter of the model, the line will be defined by an eigenvector of the $(\boldsymbol{V}^T\boldsymbol{V})^{-1}$ matrix. For nonlinear models the contours of $L(\boldsymbol{\theta})$ will not be perfect ellipses, the contours of $g(\boldsymbol{\theta})$ will not necessarily be straight lines, and the path traced by $\tilde{\boldsymbol{\theta}}$ will not be straight. However, in a region about $\hat{\boldsymbol{\theta}}$, the model and the function $g(\boldsymbol{\theta})$ will behave linearly and the path followed by $\tilde{\boldsymbol{\theta}}$ may be approximated by (6.34). Like the measures developed for ARMA models, much of the work here is based on the geometry and algebra of the linear case as developed in Section 6.4.

To overcome the computational intensity of profiling, we suggest a pseudo-profiling algorithm which eliminates the optimization problem. Instead of computing τ based on $\tilde{\boldsymbol{\theta}}$, we compute τ on the basis of values of $\boldsymbol{\theta}$ along the straight line defined by (6.34). At each iteration of the algorithm, we compute

$$\tilde{\boldsymbol{\theta}}_{approx} = \hat{\boldsymbol{\theta}} \pm \frac{s\, t(n-r, \alpha/2)(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{a}}{\sqrt{\boldsymbol{a}^T(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{a}}} \tag{6.60}$$

where $\boldsymbol{a} = \frac{\partial g}{\partial \boldsymbol{\theta}}\Big|_{\hat{\boldsymbol{\theta}}}$. Then compute

$$\tau_{approx} = \sqrt{-2ln\left(\frac{L(\tilde{\boldsymbol{\theta}}_{approx})}{L(\hat{\boldsymbol{\theta}})}\right)} \tag{6.61}$$

The resulting pseudo-profile t plot is based on the true likelihood function and therefore provides reliable information about the intrinsic nonlinearity of the solution surface in the direction defined by $(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{a}$.

We propose two plots to display the information gathered during pseudo-profiling. These are presented here for the case where the additive error is iid $N(0, \sigma^2)$. For a series of values of α, compute $\tilde{\boldsymbol{\theta}}_{approx}$. Then compute $S(\tilde{\boldsymbol{\theta}}_{approx})$ when dealing with a nonlinear regression model having iid $N(0, \sigma_a^2)$ additive random error; otherwise compute $L(\tilde{\boldsymbol{\theta}}_{appox})$. Note that

$$\frac{S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}})}{s^2} = F(1, n-r; \alpha_i) \tag{6.62}$$

for nonlinear regression models, and

$$-2\ln\left(\frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})}\right) = F(1, n-r; \alpha_i) \tag{6.63}$$

otherwise. Therefore, compute

$$F_{\tilde{\boldsymbol{\theta}}} = \frac{S(\tilde{\boldsymbol{\theta}}_{approx}) - S(\hat{\boldsymbol{\theta}})}{s^2} \tag{6.64}$$

or

$$F_{\tilde{\boldsymbol{\theta}}} = -2\ln\left(\frac{L(\tilde{\boldsymbol{\theta}})}{L(\hat{\boldsymbol{\theta}})}\right) \tag{6.65}$$

and compare this value to

$$F_{linear} = (t(n-r, \alpha_i/2))^2 \tag{6.66}$$

where F_{linear} is the value of the F statistic that would be obtained if the model and the function $g(\boldsymbol{\theta})$ were both truly linear. Then, plot $\Delta SSE = S(\tilde{\boldsymbol{\theta}}_{approx}) - S(\hat{\boldsymbol{\theta}})$ versus F_{linear}. Alternatively, or additionally, plot ΔSSE versus

$$\lambda = \sqrt{\frac{s^2 F_\alpha(1, n-r)}{\boldsymbol{a}^T(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{a}}} \tag{6.67}$$

where λ is a measure of distance from $\tilde{\boldsymbol{\theta}}_{approx}$ to $\hat{\boldsymbol{\theta}}$. On our plots we also indicate the levels of $F_{\tilde{\boldsymbol{\theta}}}$ corresponding to ΔSSE. A step-by-step algorithm for constructing a pseudo-profile plot is given in Figure 6.7. The form and interpretation of this plot is similar to one proposed by Cook and Weisberg (1990). However, they computed their results using optimized values of $\tilde{\boldsymbol{\theta}}$.

Examples of such plots are shown in Figure 6.8. These were generated for the Puromycin example of Bates and Watts (1988) and may be compared to the profile t plots for this example which are given in Figure 6.9.

Plots (a) and (c) in Figure 6.8 represent two different ways of displaying the pseudo-profiling information for θ_1, while plots (b) and (d) display the information

1. Select a value for α (we begin with $\alpha = 0.5$ and end with $\alpha = 0.05$, i.e., we begin with a 50 % confidence level and end with a 95 % confidence level).

2. Compute

$$\lambda = \sqrt{\frac{s^2 F_\alpha(1, n-p)}{\boldsymbol{a}^T(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{a}}}$$

3. Locate

$$\tilde{\boldsymbol{\theta}}_{approx} = \hat{\boldsymbol{\theta}} + \lambda(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{a}$$

4. Evaluate $S(\tilde{\boldsymbol{\theta}}_{approx})$.

5. Compute $\Delta SSE = S(\tilde{\boldsymbol{\theta}}_{approx}) - S(\hat{\boldsymbol{\theta}})$.

6. Calculate $F_{\tilde{\boldsymbol{\theta}}} = \frac{S(\tilde{\boldsymbol{\theta}}_{approx}) - S(\hat{\boldsymbol{\theta}})}{s^2}$.

7. Repeat Steps 3-6 for

$$\lambda = -\sqrt{\frac{s^2 F_\alpha(1, n-p)}{\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{a}}}$$

8. Repeat Steps 2-7 for several values of α.

9. Plot ΔSSE versus λ.

10. Plot ΔSSE versus $F_{\tilde{\boldsymbol{\theta}}}$.

Figure 6.7: A step-by-step algorithm for computing pseudo-profiles for nonlinear regression models.

for θ_2. The solid lines represent the pseudo-profiles while the dashed lines are reference lines representing the results for linear approximations to the model. These linearization reference lines are equivalent to those displayed on profile t plots, and as such, they can be used to obtain linearization confidence intervals for the parameters.

The amount by which a pseudo-profile deviates from the linearization reference line is a qualitative measure of nonlinearity. Considering the plots in Figure 6.8, the branch of the pseudo-profile for θ_1 representing values of θ_1 less than $\hat{\theta}_1$ deviates only slightly from the reference line, indicating that the nonlinearity is very small, and that the linearization result for the lower limit of the confidence interval for θ_1 would be an excellent approximation of the lower limit of a profiling-based likelihood interval for this parameter. The nonlinearity displayed for values of θ_1 greater than $\hat{\theta}_1$ is more pronounced but is still moderate, and depending on the application, the linearization approximation result may be adequate.

The pseudo-profiles for θ_2 (panels (b) and (d) in Figure 6.8) show that this parameter behaves more nonlinearly than θ_1. The dotted horizontal lines help to identify the values on the ordinate axis that correspond to various levels of confidence. These lines make it easy to see how rapidly the nonlinearity of the problem increases as the level of confidence increases (i.e. as the value of α decreases).

Although these plots do not provide a single number by which to quantify the nonlinearity, the graphical displays are easy to interpret and provide the user with much more information about the nature of the solution surface than would a single number. The plots also allow the user to decide whether or not the nonlinearity is "severe" in the context of the application.

Figure 6.9 shows the pseudo-profiling information superimposed on the two profile t plots. For this example, the pseudo-profiles are almost coincident with the true profile t plots. Only at values of τ above two do the pseudo-profiles deviate slightly from the profile t plots.

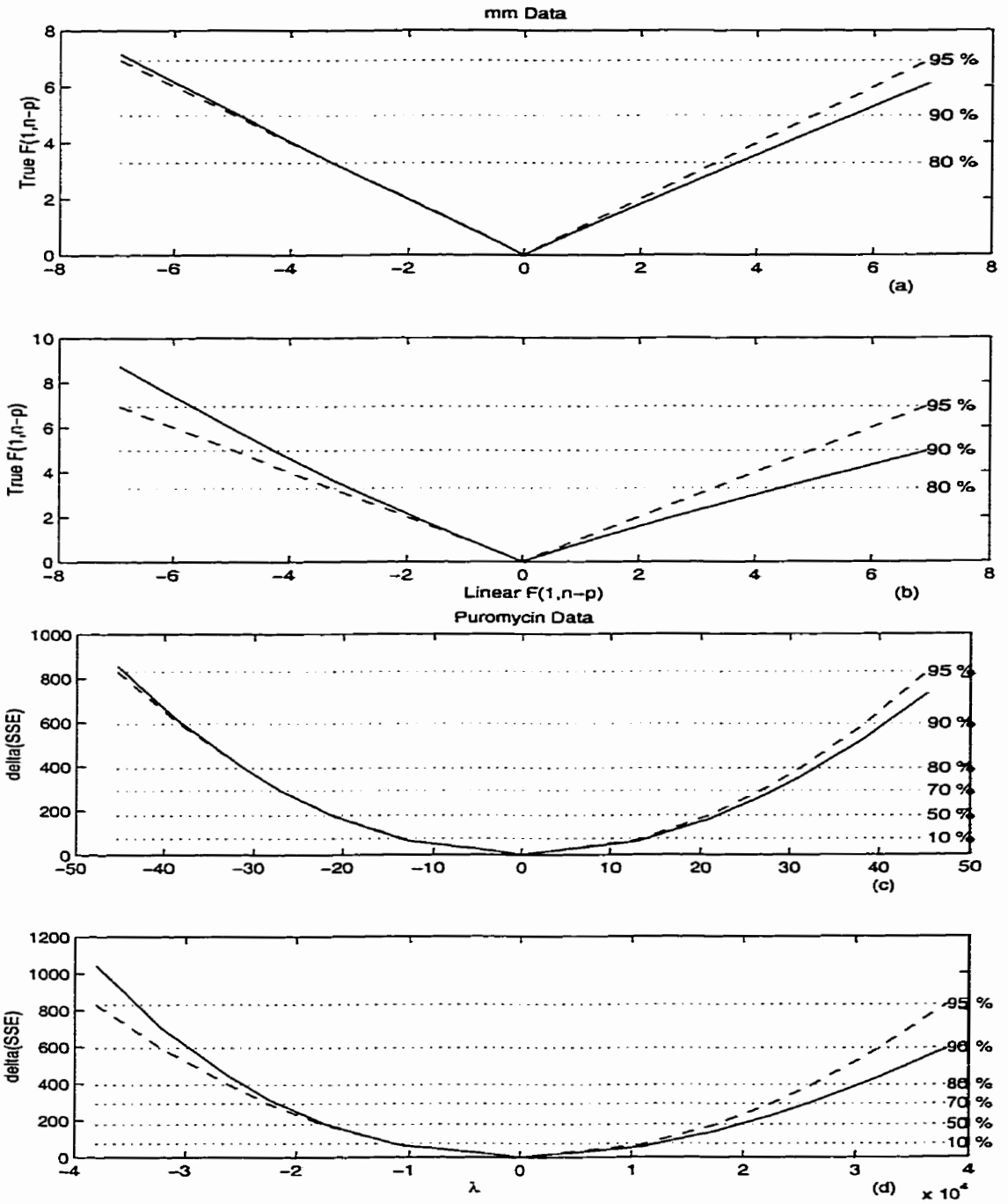

Figure 6.8: Pseudo-Profile plots for the parameters of the Puromycin Example. KEY: – pseudo-profile plots; - - reference line.

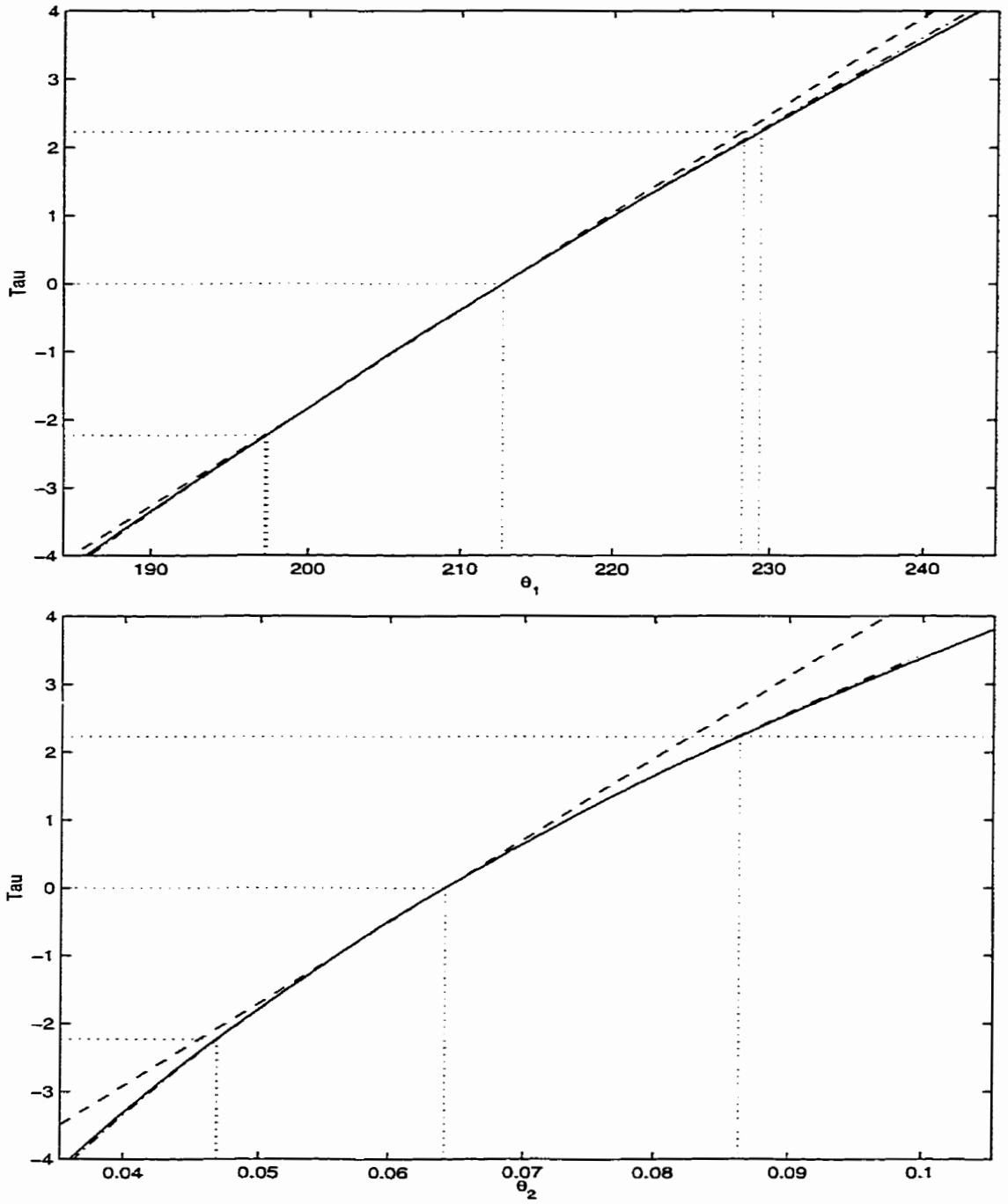

Figure 6.9: Profile t plots and pseudo-profiles for the parameters of the Puromycin Example. KEY: – profile t plots; .- pseudo-profiles; - - reference line.

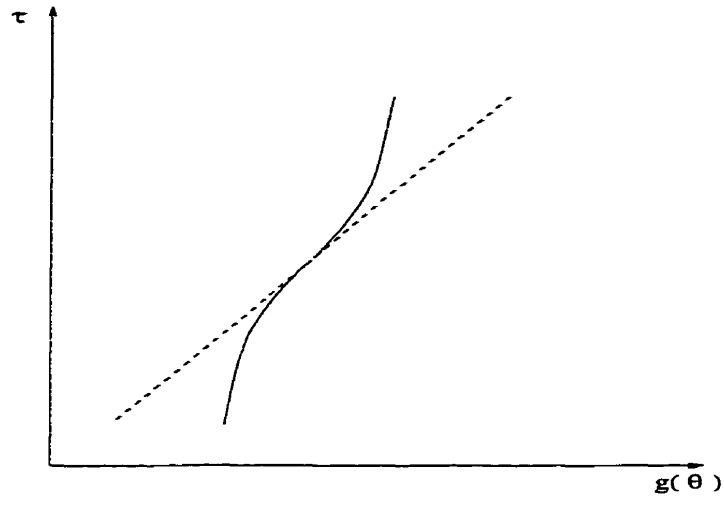

Figure 6.10: A sketch of a profile t plot for which the likelihood interval is shorter than the linearization interval.

It is possible to predict the direction in which a pseudo-profile will deviate from the corresponding profile-t plot. When the true profile t plot shows nonlinearity of the form shown in the sketch in Figure 6.10 (i.e. when the likelihood interval is narrower than the linearization confidence interval), pseudo-profiles overestimate the nonlinearity because they are not based on optimized values of $\tilde{\boldsymbol{\theta}}$. At any point on the pseudo-profile, $S(\tilde{\boldsymbol{\theta}}_{approx})$ will always have a value greater than or equal to $S(\tilde{\boldsymbol{\theta}})$, where $S(\tilde{\boldsymbol{\theta}})$ is the sum of squares of residuals used to construct the true profile t plots. In this way, the pseudo-profile will lie on or above the profile t plot for values of $g(\boldsymbol{\theta})$ greater than $g(\hat{\boldsymbol{\theta}})$, and on or below the profile t curve for values of $g(\boldsymbol{\theta})$ less than $g(\hat{\boldsymbol{\theta}})$. Conversely, when the true profile displays nonlinearity that has the general shape shown in Figure 6.11, the pseudo-profile will tend to underestimate the nonlinearity because $S(\tilde{\boldsymbol{\theta}}_{approx})$ will again be equal to or greater than $S(\tilde{\boldsymbol{\theta}})$, but in this case the increase in the sum of squares of residuals will tend to place the pseudo-profile closer to the the linearization confidence interval. However, we have found in practice that pseudo-profiles are good approximations to the true profile t plots and that they capture the nonlinearity of the problem.

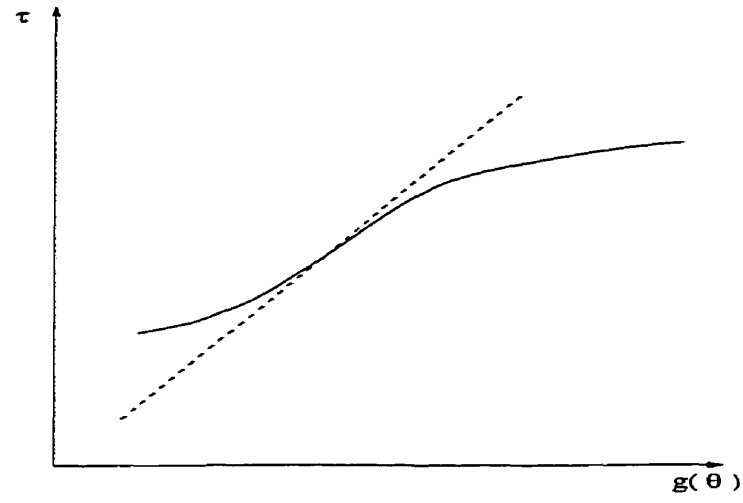

Figure 6.11: A sketch of an 'S' type profile t plot.

6.7 Conclusions

A new measure of nonlinearity for time series models has been proposed. For the examples considered here, and in other cases we have studied, the indications of nonlinearity provided by the measure ζ_{min} correlates well with the observed nonlinearities. However, more work is required to see how ζ_{min} performs for a broader range of examples. We have found that significant nonlinearity can be expected when ζ_{min} is less than one. As a measure of nonlinearity, ζ_{min} has the added advantage that its expected value can be computed in the absence of data. $\zeta_{min,ap}$ may be used to provide important information about how much nonlinearity to expect in the results of a proposed application. This could influence the amount of data collected, and ultimately, the statistical tools required to analyze the data.

For the case of nonlinear regression models, a pseudo-profiling algorithm has been proposed for judging qualitatively the amount of nonlinearity to expect in a function $g(\boldsymbol{\theta})$. This algorithm is more attractive than the true profiling algorithm because it does not involve any numerical optimizations. It also does not require the computation of Hessian matrices as do the measures of nonlinearity previously proposed in the literature. Although this methodology has been illustrated and discussed in the

context of nonlinear regression models, it can also be used with any class of models as long as a likelihood function for the parameters of the model can be specified.

The two approaches to measuring nonlinearity proposed in this paper are meant to be "quick and easy" indicators of the degree of nonlinearity of individual parameters and functions of parameters of proposed models. They are meant to be reliable, yet it avoid the complexity and computational intensity of measures of nonlinearity proposed to date.

6.8 Acknowledgements

The authors gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada and the School of Graduate Studies of Queen's University.

6.9 Nomenclature

$\boldsymbol{a}$	$=p \times 1$ vector of partial derivatives of $g(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$
b	= a constant
c	= a constant
$\boldsymbol{e}$	$= n \times 1$ column vector of residuals
$F(p, n-p; \alpha)$	= upper α quantile for the F distribution with p and $n-p$ degrees of freedom
$g(\boldsymbol{\theta})$	= a function of parameters
$\boldsymbol{h}$	= any nonzero $p \times 1$ vector
$\boldsymbol{H}$	$= n \times p \times q$ Hessian array of second derivatives of $f(\boldsymbol{x}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$

$\boldsymbol{k}$	$= p \times 1$ vector of ones and minus ones given by $sign(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_b)$
ℓ	= a measure of the distance from $\hat{\boldsymbol{\phi}}$ to the nearest stability/invertibility boundary
$L(\boldsymbol{\theta})$	= likelihood function evaluated at $\boldsymbol{\theta}$
m_i^{ϕ}	= marginal curvature for ϕ_i
n	= number of observations
p	= order of an AR polynomial
q	= order of an MA polynomial
r	= total number of estimated parameters
s	= estimated standard deviation of the random errors
$S(\boldsymbol{\theta})$	= sum of squared errors
se	= standard error
$t(n-p;\alpha/2)$	= upper $\alpha/2$ quantile of the t distribution with $n-p$ degrees of freedom
$\boldsymbol{V}$	$= n \times p$ matrix of elements v_{ij} representing the first derivative of $f(\boldsymbol{x}_i, \boldsymbol{\theta})$ with respect to the j^{th} parameter
$\boldsymbol{x}$	$= 1 \times m$ row vector of m independent variables
$\boldsymbol{X}$	$= n \times p$ matrix of elements x_{ij} representing the level of the j^{th} independent variable for observation i
y	= response variable
$\boldsymbol{y}$	$= n \times 1$ column vector of values of the response variable

Greek letters

α = significance level

γ_h^N = intrinsic nonlinearity

γ_h^T = parameter-effects curvature

γ_{max}^N = maximum intrinsic curvature

γ_{max}^T = maximum parameter-effects curvature

$(\gamma_{RMS}^N)^2$ = root mean square intrinsic curvature

$(\gamma_{RMS}^T)^2$ = root mean square parameter-effects curvature

$\delta(\theta_i)$ = studentized value of θ_i

$\delta(g(\boldsymbol{\theta}))$ = studentized value of $g(\boldsymbol{\theta})$

ϵ = additive random error

$\boldsymbol{\epsilon}$ = $n \times 1$ column vector of random errors

ζ = measure of nonlinearity for ARMA models

ζ_{min} = minimum value of the $2p$ values of ζ

$\dot{\eta}_h$ = tangent vector to the solution surface at $b = 0$

$\ddot{\eta}_h$ = acceleration of an arbitrary straight line in the space of the parameters

$\ddot{\eta}_h^N$ = acceleration normal to the tangent line

$\ddot{\eta}_h^P$ = acceleration parallel to $\dot{\eta}_h$

$\ddot{\eta}_h^G$ = acceleration parallel to the tangent plane and normal to $\dot{\eta}_h$

θ_i = i^{th} parameter of a model

$\theta(B)$ = moving average polynomial of a time series model

$\boldsymbol{\theta}$ = $p \times 1$ column vector of parameters of a model

$\boldsymbol{\theta}_0$ = point in the parameter space

$\bar{\boldsymbol{\theta}}_{approx}$ = location of a limit of a linearization confidence interval

λ = constant which defines the confidence level

$\nu(u)$ = covariance at lag u

ξ = ratio of the distance from $\hat{\boldsymbol{\phi}}$ to $\boldsymbol{\phi}_b$ to the distance from $\hat{\boldsymbol{\phi}}$ to $\boldsymbol{\phi}_{95}$

$\rho(u)$ = correlation at lag u

σ_a^2 = variance of the white noise sequence a_t

$\tau(\theta_i)$ = profile t statistic for θ_i

$\tau(g(\boldsymbol{\theta}))$ = profile t statistic for $g(\boldsymbol{\theta})$

Υ = expected value of $\boldsymbol{\mathcal{I}}$

$\boldsymbol{\phi}$ = $p \times 1$ vector of parameters defined by the partial autocorrelation transformation

$\boldsymbol{\phi}_{b,i,j}$ = point where the vector joining $\hat{\boldsymbol{\phi}}$ and $\boldsymbol{\phi}_{95,i,j}$ crosses a stability/invertibility boundary

$\boldsymbol{\phi}_{95,i,1}$ = upper limit of the approximate 95 % confidence interval for ϕ_i

$\boldsymbol{\phi}_{95,i,2}$ = lower limit of the approximate 95 % confidence interval for ϕ_i

$\phi(B)$ = autoregressive polynomial of a time series model

$\chi^2(1)$ = the chi-squared distribution with 1 degree of freedom

$\psi(t, \boldsymbol{\theta})$ = $p \times 1$ vector of derivatives of a_t with respect to $\boldsymbol{\theta}$

$\sigma_a^2 \boldsymbol{\Omega}_n$ = $n \times n$ covariance matrix for $\boldsymbol{y_n}$

Superscripts

$*$ = a true value

$\hat{}$ = a maximum likelihood estimate

$\tilde{}$ = a constrained estimate

Abbreviations

AR	autoregressive
ARMA	autoregressive moving average
iid	independently and identically distributed
MA	moving average
MLE	maximum likelihood estimate
PACF	partial autocorrelation function

Chapter 7

Measuring Uncertainty in Control-Relevant Statistics

7.1 Abstract

To make appropriate decisions based on common indices used in control, both the point estimates and their uncertainties must be known. Furthermore, the burgeoning field of robust control requires that bounds on the model uncertainty be specified. We present generalized profiling as a means by which to reliably estimate the parametric uncertainty in functions of parameters of transfer function models. Many control-relevant statistics, such as model predictions and gain margins, can be conceived of as being functions of parameters of proposed process models. Generalized profiling is a flexible tool for measuring uncertainty due to estimated parameters, and therefore, can satisfy the demands of robust control. For multi-step-ahead predictions, we develop a methodology for computing likelihood intervals which accounts for both parameter uncertainty and uncertainty due to unknown future noise.

When more than one function of parameters is being considered simultaneously, it is appropriate to quantify the joint uncertainty by specifying a likelihood region.

We adapt the profile pair algorithm of Bates and Watts (1988) to efficiently sketch joint likelihood regions for two or more functions of parameters. To illustrate the method, likelihood regions are sketched around several points on Nyquist plots. The likelihood approach preserves the important engineering characteristics of the problem, including the fact that the uncertainty in the imaginary component goes to zero at steady state.

7.2 Introduction

In the design and tuning of controllers, there is almost always a need for a process model. Early approaches to control assumed that the model described the process exactly; fine tuning of the controller after design and implementation was used to adjust for any discrepancies between the model and the true plant. Later, those in the field of robust control embraced the premise that all models are wrong, and a plethora of design strategies were put forward which took account of known model uncertainty at the design stage (Morari and Zafiriou, 1989). Despite the incompatibility of the notion that a model can not be known exactly, with the notion that the limits on the uncertainty can be specified exactly, those in the field of system identification responded to the demands of robust control and proposed algorithms for identifying hard error bounds for process models (Gunnarson, 1993; Zhu, 1989). Hard error bounds are guaranteed upper bounds on model uncertainty. This concept is at odds with traditional statistical uncertainty (confidence) bounds which have a specific probability of enclosing the true value of the statistic. Statistical uncertainty bounds, also called soft error bounds, have been developed in the area of system identification, and these now seem to be in favor (Goodwin et al., 1992; Schoukens and Pintelon, 1994; DeVries and Van den Hof, 1995, Canale et al., 1998). In this paper we present in detail a statistical algorithm for estimating uncertainty in model parame-

ters, model predictions, and any other function of the parameters. We suggest that this "soft" approach to quantifying uncertainty is consistent with sound statistical practice, and indeed, the philosophy underlying the field of robust control.

There are three main factors which contribute to uncertainty in models fitted to data:

- noise in the data
- changing plant dynamics
- a choice of model form which can not capture the true process dynamics (Zhou and Kimura, 1994).

In this work, we assume that plant dynamics are constant, and that the "true" process can be adequately described by a model of the proposed form. Still, the noise inherent in all experimental data causes models fitted to data to be approximate. They are approximate in that there is uncertainty in the estimates of the parameters. This uncertainty in the parameters propagates to any other quantity calculated on the basis of the proposed model. How "good" a model is deemed to be depends on the purpose for which the model is being used. For example, it may be that the estimates of the parameters of a proposed model are highly uncertain, but this same model may be able to give predictions which are close to the true values. In that case, the model would be "bad" if the values of the parameters themselves were of interest, but would be "good" if the model were to be used for predictive purposes.

The field of robust control has had a positive influence on system identification in that it has brought to the forefront issues about model uncertainty. However, it has also meant that most of the attention has been focused on uncertainty in the model predictions or frequency response. There are many other quantities (control-relevant statistics) which are calculated and used routinely in control. Many of these can be looked upon as being functions of the parameters of the proposed model. For example,

the gain margin is an important control-relevant statistic which is a function of the parameters of a proposed model. The gain margin can be defined as the inverse of the amplitude ratio at the crossover frequency, where the crossover frequency is any frequency at which the phase shift is -180^o. The gain margin, then, depends only on the form of the model and the values of its parameters. If the parameters of the model are only estimates of the true parameters, and are therefore uncertain, then the gain margin, or any other function of the parameters, calculated on the basis of the proposed model, is also uncertain. Other examples of functions of parameters used in process control include: model predictions, crossover frequencies, and many measures of controller performance (Shirt, 1997). Generalized profiling is presented here as a means of computing reliable likelihood intervals for functions of parameters.

The importance of generalized profiling extends beyond the realm of robust control. In all aspects of engineering it is as important to quantify the uncertainty in an estimated value as it is to quantify the value itself. Typically, those who have recognized the need for uncertainty estimates have reported the uncertainties as confidence intervals based on linear approximations to the model and to the function of parameters of interest. These linearization confidence intervals have been shown to be often unreliable, and sometimes misleading when used in conjunction with regression models (Donaldson and Schnabel, 1987). In this paper we discuss nonlinearity observed in control-relevant statistics, and we compare linearization confidence intervals to likelihood intervals obtained using generalized profiling. Comments are made about the adequacy of the linearization results in the context of functions of parameters of transfer function models. We demonstrate the application of profiling to individual functions of parameters, including the gain margin and multi-step-ahead predictions, and to multiple functions of parameters considered simultaneously, such as points on a Nyquist plot. Shirt (1997) examined generalized profiling for estimating uncertainty in measures of controller performance. The profiling algorithm failed

in this case (Quinn et al., 1999a; Chapter 3), and an alternate method was proposed. We also discuss issues related to the use of profiling in cases where a fitting criterion other than maximum likelihood is used.

This paper is structured as follows. In this section, some examples of functions of parameters of interest in control have been introduced; subsequently, these examples serve illustrative and investigative purposes throughout the paper. In Section 7.3, generalized profiling is introduced along with some alternative approaches, including the linearization approach. Section 7.4 provides details regarding computational issues. Section 7.6 contains several examples used to illustrate generalized profiling, and to identify its merits and limitations. The focus is on demonstrating the use of generalized profiling in control applications, and on connecting generalized profiling to issues in system identification, dynamic design of experiments, and robust control.

7.3 Uncertainty Intervals

7.3.1 The Linearization Approach to Confidence Intervals

In this paper the term "linear model" is being used in the statistical sense to mean a model which is linear in the parameters. Accordingly, the term "nonlinear model" used in the statistical sense should be distinguished from the corresponding term used in the engineering sense, which refers to models which are nonlinear in the state variables or regressor variables.

For a single input single output (SISO) model, a general discrete transfer function model can be written as (Ljung, 1987)

$$A(q^{-1})y_t = \frac{B(q^{-1})}{F(q^{-1})}q^{-d}u_t + \frac{C(q^{-1})}{D(q^{-1})}a_t \tag{7.1}$$

where q^{-1} is a backshift operator such that $q^{-1}y_t = y_{t-1}$ and $A(q^{-1})$, $B(q^{-1})$, $C(q^{-1})$,

$D(q^{-1})$ and $F(q^{-1})$ are polynomials having the form:

$$A(q^{-1}) = 1 + a_1 q^{-1} + \ldots + a_{na} q^{-na} \tag{7.2}$$

$$B(q^{-1}) = b_0 + b_1 q^{-1} + \ldots + b_{nb} q^{-nb} \tag{7.3}$$

$$C(q^{-1}) = 1 + c_1 q^{-1} + \ldots + c_{nc} q^{-nc} \tag{7.4}$$

$$D(q^{-1}) = 1 + d_1 q^{-1} + \ldots + d_{nd} q^{-nd} \tag{7.5}$$

$$F(q^{-1}) = 1 + f_1 q^{-1} + \ldots + f_{nf} q^{-nf} \tag{7.6}$$

and d is the delay between a change in the manipulated input variable u_t and its effect on the output variable y_t. Note that this general model includes within its structure specific model types such as the autoregressive moving average (ARMA) model, the autoregressive moving average behavior with exogenous inputs (ARMAX) model and the Box-Jenkins model (Ljung, 1987). It can be verified that a general transfer function model is nonlinear in the parameters. That is, the partial derivative of y_t with respect to at least one of the parameters is, itself, a function of the parameters. The exception is the ARX model (ARMAX model with no moving average component) which is linear in the parameters since it can be written as:

$$A(q^{-1}) y_t = B(q^{-1}) q^{-d} u_t + a_t \tag{7.7}$$

and in difference form as:

$$y_t = b_0 + b_1 u_{t-k-1} + \ldots + b_{nb} u_{t-k-nb} - a_1 y_{t-1} - \ldots - a_{na} y_{t-na} + a_t \tag{7.8}$$

From the difference equation we can see that y_t is a linear function of past inputs and outputs with additive normal white noise. It has been shown that linear least squares inference results are exact for this case (Söderström and Stoica, 1989). Note that a finite impulse response (FIR) model is a linear model included within the ARX structure.

Although confidence intervals are widely quoted in the system identification literature, it is important to understand that they are almost invariably approximate confidence intervals based on first linearizing the nonlinear model and any nonlinear function of the parameters. This linear approximation approach is attractive in that it provides computationally simple results. However, Bates and Watts (1980), Donaldson and Schnabel (1987), and others have shown that, in the case of nonlinear regression models, linear approximations are often unreliable and possibly misleading.

For all types of models (e.g. mechanistic, empirical, steady-state, dynamic, etc.) which are nonlinear in the parameters it is common to estimate confidence intervals for individual parameters by first taking a linear approximation to the nonlinear model and then applying the linearization inference results. The following is an overview of the linearization approach to confidence intervals, based on Quinn et al. (1999a) (Chapter 3).

Consider a general model of the form:

$$y_t = f(\boldsymbol{x_t}, \boldsymbol{\theta}) + a_t \tag{7.9}$$

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 $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$ at time t, $\boldsymbol{\theta}$ is

a vector of p parameters, and a_t is the additive random error term associated with y_t. The terminology used to refer to this model and its variables changes depending on the context. The independent variables $\boldsymbol{x}$ are also referred to as input variables or manipulated variables. y_t is also called an output variable or a controlled variable. Equation 7.9 can be written in terms of a vector of n observed response values $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$:

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}) + \boldsymbol{a} \tag{7.10}$$

where $\boldsymbol{X}$ is an $n \times m$ matrix of $\boldsymbol{x}$ values or functions of $\boldsymbol{x}$ with element $x_{i,j}$ representing the level of the j^{th} $\boldsymbol{x}$ variable for observation i, and $\boldsymbol{a}$ is the $n \times 1$ vector of a_t values.

A linear model is a special case of (7.9) where

$$\boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}) = \boldsymbol{X}\boldsymbol{\theta} \tag{7.11}$$

and $\boldsymbol{X}$ is an $n \times p$ matrix of functions of the $\boldsymbol{x}$ values. When the model is linear and the noise is additive and identically and independently normally distributed (i.e., $iid\, N(0, \sigma^2)$), then

$$y_t = \boldsymbol{x}^T\boldsymbol{\theta} + a_t \tag{7.12}$$

where $\boldsymbol{x}^T$ is a p-element row vector of functions of the $\boldsymbol{x}$ values, and the exact analytic confidence interval for the i^{th} parameter is:

$$\hat{\theta}_i \pm se(\hat{\theta}_i)\, t(n-p, \alpha/2) \tag{7.13}$$

where $\hat{\theta}_i$ is the least squares estimate of the i^{th} parameter, $t(n-p, \alpha/2)$ is the upper

$\alpha/2$ quantile for the t distribution with $n-p$ degrees of freedom,

$$se(\hat{\theta}_i) = s\sqrt{[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]_{ii}} \tag{7.14}$$

is the standard error of the i^{th} parameter estimates, $s = \sqrt{\frac{S(\hat{\boldsymbol{\theta}})}{n-p}}$ is an estimate of the standard deviation of the random error: $S(\hat{\boldsymbol{\theta}}) = \boldsymbol{e}^T\boldsymbol{e}$ is the minimum sum of squared residuals for the fitted model, $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ is the vector of residuals, $\hat{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{X}, \hat{\boldsymbol{\theta}})$ is a vector of fitted response values, and $[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]_{ii}$ is the i^{th} diagonal entry in the inverse of the information matrix.

A $(1-\alpha)100\%$ confidence interval for a parameter is the set of values of that parameter which has a probability of at least $(1-\alpha)100\%$ of including the true value of the parameter. The confidence interval for any linear function of the parameters $g(\boldsymbol{\theta}) = \boldsymbol{h}^T\boldsymbol{\theta}$ is:

$$g(\hat{\boldsymbol{\theta}}) \pm s\, t(n-p;\alpha/2)\sqrt{\boldsymbol{h}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{h}} \tag{7.15}$$

where $\boldsymbol{h}$ is a $p \times 1$ vector of constants. Note that the expected value of a prediction at $\boldsymbol{x}_k$ from a linear model is a linear function of the parameters where $\boldsymbol{h}^T\boldsymbol{\theta} = \boldsymbol{x}_k\boldsymbol{\theta}$.

There are no such exact analytic results for the confidence intervals when the model is nonlinear. To use the linearization results in the case of a general nonlinear model, the nonlinear model and the function of parameters $g(\boldsymbol{\theta})$ can each be linearized using first order Taylor Series approximations centered at $\hat{\boldsymbol{\theta}}$. Then, the well-known linear inference results can be applied. That is, the likelihood interval for any function $g(\boldsymbol{\theta})$ is

$$g(\hat{\boldsymbol{\theta}}) \pm s\, t(n-p;\alpha/2)\sqrt{\frac{\partial g}{\partial \boldsymbol{\theta}}(\boldsymbol{V}^T\boldsymbol{V})^{-1}\frac{\partial g}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}} \tag{7.16}$$

where $\boldsymbol{V}$ is an $n \times p$ matrix with elements defined by $\boldsymbol{V}_{ij} = \frac{\partial f(x_i,\theta)}{\partial \theta_j}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$. Note that

the values of some of the elements of $\boldsymbol{V}$ depend on the initial conditions of the $\{a_t\}$ series. We exploit the fact that

$$\frac{\partial y_t}{\partial \boldsymbol{\theta}} = -\frac{\partial a_t}{\partial \boldsymbol{\theta}} \tag{7.17}$$

to compute the derivatives numerically using the prediction errors obtained using Kalman filtering (see Section 7.4). In this way, we appropriately account for the unknown initial conditions in the calculation of $\boldsymbol{V}$. Is also possible to estimate the initial conditions using backcasting (Box and Jenkins, 1976) and to then use these estimates to compute the elements of $\boldsymbol{V}$.

The linear approximation to the general transfer function is

$$f(\boldsymbol{x}, \boldsymbol{\theta}) \approx \hat{y}_{t|t-1}(\hat{\boldsymbol{\theta}}) + \left.\frac{\partial \hat{y}_{t|t-1}}{\partial \boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \tag{7.18}$$

where

$$\hat{y}_{t|t-1}(\boldsymbol{\theta}) = \frac{G(q^{-1})}{H(q^{-1})} q^{-d} u_t + \left[1 - \frac{1}{H(q^{-1})}\right] y_t \tag{7.19}$$

and $\boldsymbol{\theta}^T = (a_1, \ldots, a_{na}, b_0, b_1, \ldots, b_{nb}, c_1, \ldots, c_{nc}, d_1, \ldots, d_{nd}, f_1, \ldots, f_{nf})$ (i.e., $\boldsymbol{\theta}$ is the vector of all parameters to be estimated) with

$$G(q^{-1}) = \frac{B(q^{-1})}{A(q^{-1})F(q^{-1})} q^{-d} \tag{7.20}$$

and

$$H(q^{-1}) = \frac{C(q^{-1})}{A(q^{-1})D(q^{-1})} \tag{7.21}$$

Linear inference results for any nonlinear function of the parameters, $g(\boldsymbol{\theta})$, can be

obtained by linearizing the function as

$$g(\boldsymbol{\theta}) = g(\hat{\boldsymbol{\theta}}) + \frac{\partial g}{\partial \boldsymbol{\theta}}\Big|_{\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \tag{7.22}$$

In system identification, it is often argued that such a linearization approach is justified on the basis that the linear inference results are the asymptotic result as $n \to \infty$, and large amounts of data are common. However, there has been little evidence to support these claims, and the amount of data which constitutes a "large amount" is seldom quantified.

Although the linearization approach provides a quick analytic means of finding confidence intervals, it can be misleading because it does not account for the curvature of the expectation surface nor the nonlinearity of the mapping of $\boldsymbol{\theta}$ from the observation space to the parameter space (Donaldson and Schnabel, 1987; Ratkowsky, 1983; Bates and Watts, 1980). Taking account of these shortcomings is the motivation for profiling, which, being an optimization-based numerical method, is more computationally intensive.

7.3.2 Profiling

Profiling (Bates and Watts, 1988; Chen, 1991; Lam and Watts, 1991; Severini and Staniswalis, 1994; Chen and Jennrich, 1996, Quinn et al., 1999a; Chapter 3) is a graphical means by which to display inference results for parameters, and functions of parameters, of proposed models. Bates and Watts (1988) developed the algorithm specifically to summarize inferential results for parameters of nonlinear regression models. Lam and Watts (1991) extended the theory of profiling to encompass time series models using a modified sum of squares appropriate for time series models. Chen (1991), and Chen and Jennrich (1996) developed the theory of profiling in terms of likelihood ratios and constrained optimization. This formulation of the

profiling algorithm is very general and may be used with several classes of models, including time series models. Furthermore, the constrained optimization approach readily accommodates the computation of inference results for functions of parameters (Quinn et al., 1999a; Chapter 3).

Profiling is based on the τ statistic

$$\tau(g(\boldsymbol{\theta})) = sign(g(\boldsymbol{\theta}) - g(\hat{\boldsymbol{\theta}}))\sqrt{-2ln\left(\frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})}\right)} \tag{7.23}$$

This expression for τ 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 can be found.

It is common to express uncertainty in an individual estimated value by a confidence interval. Although often abused, the term "confidence interval" has a precise statistical definition rooted in the frequency theory of probability (Kendall and Stuart, 1967). We will use the term "likelihood interval" to describe an uncertainty interval obtained using generalized profiling so as to respect the statistical definition of confidence interval. A nominal $(1-\alpha)100\%$ likelihood interval for $g(\boldsymbol{\theta})$ is the set of all values of $g(\boldsymbol{\theta})$ which are plausible with probability $(1-\alpha)100\%$ given the available data. From standard asymptotic arguments (Cox and Hinkley, 1974),

$$LI\left(g(\boldsymbol{\theta})\right) = \left\{g(\boldsymbol{\theta}) : -2\ln\left(\frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})}\right) \leq \chi^2_\alpha(1)\right\} \tag{7.24}$$

where $LI(g(\boldsymbol{\theta}))$ is the $(1-\alpha)100\%$ likelihood interval for $g(\boldsymbol{\theta})$, $L(\boldsymbol{\theta})$ is the likelihood function for $\boldsymbol{\theta}$, and $\chi^2_\alpha(1)$ is the upper α quantile for the χ^2 distribution with 1 degree of freedom, $\hat{\boldsymbol{\theta}}$ is the vector of maximum likelihood estimates of the parameters, and $\boldsymbol{\theta}$ is any allowable vector of parameter values (Chen and Jennrich, 1996). Equation (7.24) states that a $(1-\alpha)100\%$ likelihood interval for $g(\boldsymbol{\theta})$ includes all possible values

of $g(\boldsymbol{\theta})$ over the region in the parameter space defined by $-2\ln\left(\frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})}\right) \leq \chi^2_\alpha(1)$. Note that the likelihood ratio statistic follows *asymptotically* the $\chi^2(1)$ distribution (Cordeiro et al., 1994), except in special cases where it is exact. In order to account for uncertainty in the value of the variance of the random error σ^2, we compute $(1-\alpha)100\%$ likelihood intervals based on

$$LI\left(g(\boldsymbol{\theta})\right) = \left\{ g(\boldsymbol{\theta}) : -2\ln\left(\frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})}\right) \leq F(1, n-p; \alpha) \right\} \tag{7.25}$$

where $F(1, n-p; \alpha/2)$ is the upper α quantile for the F distribution with 1 and $n-p$ degrees of freedom (Cook and Weisberg, 1990). In other words, a $(1-\alpha)100\%$ likelihood interval for $g(\boldsymbol{\theta})$ is the set of all $g(\boldsymbol{\theta})$ for which

$$-t(n-p, \alpha/2) \leq \tau(g(\boldsymbol{\theta})) \leq t(n-p, \alpha/2) \tag{7.26}$$

where $t(n-p, \alpha/2)$ is the upper $\alpha/2$ quantile for the student t distribution with $n-p$ degrees of freedom (Bates and Watts, 1988).

To find a likelihood interval for $g(\boldsymbol{\theta})$, we solve a series of constrained optimization problems of the form:

Maximize:

$$L(\boldsymbol{\theta}) \tag{7.27}$$

subject to:

$$g(\boldsymbol{\theta}) = c$$

Let the location of the solution to this problem be $\tilde{\boldsymbol{\theta}}$. Then, a profile t plot is a plot of $\tau(g(\tilde{\boldsymbol{\theta}}))$ versus $g(\tilde{\boldsymbol{\theta}}))$ for a range of values of c. The limits of a $(1-\alpha)100\%$ likelihood

interval for $g(\boldsymbol{\theta})$ can be read from the profile t plot by finding those values of $g(\boldsymbol{\theta})$ which define the points on the profile at $\tau = \pm t(n-p, \alpha/2)$. 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 inference results would be. A reference line, $\delta(g(\boldsymbol{\theta}))$ versus $g(\boldsymbol{\theta})$, is typically included on profile t plots, where

$$\delta(g(\boldsymbol{\theta})) = \frac{g(\boldsymbol{\theta}) - g(\hat{\boldsymbol{\theta}})}{se(g(\hat{\boldsymbol{\theta}}))} \tag{7.28}$$

This reference line may be used to obtain the linearization confidence intervals for $g(\boldsymbol{\theta})$, and to judge the relative curvature of the parameter or function or parameters (Chen, 1991). The generalized profiling algorithm is given in Figure 7.1.

7.3.3 Other Approaches to Estimating Uncertainty

As an alternative to the graphical profiling approach, the upper limit of a likelihood interval for $g(\boldsymbol{\theta})$ may be found by reformulating the optimization problem as follows:

$$\begin{aligned} &\text{Maximize} \quad g(\boldsymbol{\theta}) \\ &\text{Subject to} \quad S(\boldsymbol{\theta}) \leq \gamma \end{aligned} \tag{7.29}$$

where γ is a constant chosen based on the desired level of confidence, and the variance of the additive white noise. In this work we use

$$\gamma = S(\hat{\boldsymbol{\theta}}) \left[1 + \frac{1}{n-p} F(1, n-p; \alpha) \right] \tag{7.30}$$

Similarly, the lower limit of the likelihood interval may be found by minimizing $g(\boldsymbol{\theta})$ subject to the same constraint. This procedure amounts to locating the maximum and minimum of the function $g(\boldsymbol{\theta})$ over the joint confidence region for $\boldsymbol{\theta}$. Chen (1991), and Khorasani and Milliken (1982) proposed approximations of this minimiza-

1. Using a nonlinear optimization package, find the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$.

2. Compute the MLE of $g(\boldsymbol{\theta})$, and define $\hat{g} = g(\hat{\boldsymbol{\theta}})$.

3. Compute an estimate of the variance of the additive random error (i.e., compute s^2).

4. Compute $Cov(\hat{\boldsymbol{\theta}})$, the covariance of the estimated parameters.

5. Compute $se(\hat{g}) = \sqrt{s^2 \left.\frac{dg}{d\boldsymbol{\theta}}^T\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{V}^T\boldsymbol{V})^{-1} \left.\frac{dg}{d\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}}$.

6. Set the index i to 1, and let $g_{old} = \hat{g}$.

7. Move the value of $g(\boldsymbol{\theta})$ away from $\hat{g}$ by a small amount Δ (i.e., $g_i = g_{old} + \Delta$). A good starting value for Δ is $se(\hat{g})/5$.

8. Use a constrained nonlinear optimization package to solve the constrained optimization problem, maximize $L(\boldsymbol{\theta})$ subject to $g(\boldsymbol{\theta}) = g_i$. The location of the constrained optimum is $\tilde{\boldsymbol{\theta}}$.

9. Compute

$$\begin{aligned} \tau_i &= sign(g_i - \hat{g})\sqrt{-2ln\left(\frac{L(\tilde{\boldsymbol{\theta}})}{L(\hat{\boldsymbol{\theta}})}\right)} \\ \delta_i &= \frac{g_i - \hat{g}}{se(\hat{g})} \end{aligned}$$

10. Is $|\tau_i| \geq t(n-p, \alpha/2)$? If yes, continue. If no, set $g_{old} = g_i$, set $i = i + 1$, and return to Step 7.

11. Is Δ negative? If yes, continue. If no, set $g_{old} = \hat{g}$, set $i = i + 1$, let $\Delta = -\Delta$ and return to Step 7.

12. Fit a smooth curve through g_i versus τ_i and use this to find the values of $g(\boldsymbol{\theta})$ at $\tau = \pm t(n-p, \alpha/2)$. These are the limits of the $(1-\alpha)100\%$ likelihood interval.

13. Compute the limits of the $(1-\alpha)100\%$ linearization confidence interval using

$$CI = \hat{g} \pm se(\hat{g})t(n-p, \alpha/2)$$

14. Construct the profile t plot by plotting, on one figure, τ_i versus g_i, and δ_i versus g_i.

Figure 7.1: A step-by-step algorithm for profiling a function of parameters $g(\boldsymbol{\theta})$.

tion/maximization procedure. Although the minimization/maximization procedure results in the same uncertainty intervals as those obtained by profiling, profiling is considered superior because:

1. the optimization problem solved during each iteration of the profiling algorithm is more likely to converge than the optimization problem in (7.29).

2. the graphical presentation of the information gathered when profiling is useful for assessing the overall behavior of the solution surface.

Cook and Weisberg (1990) and Clarke (1987) proposed methods for estimating uncertainty which are closely related to profiling.

Resampling methods have been used to obtain likelihood intervals for parameters of nonlinear models (Alpen and Gelb 1990; Bolviken and Skovlund, 1996). These methods are based on finding empirical approximations to the distribution of the statistic of interest by repeatedly simulating the system and estimating the statistic. Such methods include: Monte Carlo simulation, the Jackknife procedure, and the Bootstrap procedure. These methods are attractive because they require no *a priori* knowledge or assumptions about the distribution of the statistic.

The demands of robust control theory have motivated recent work on quantifying uncertainty in proposed process models. Common hard bounding algorithms include: the "unknown-but-bounded noise" algorithm, the "ellipsoidal" algorithm and "set membership" algorithms (Goodwin et al., 1992). Wahlberg and Ljung (1992) used set membership theory with a geometrical justification to develop hard error bounds for linear transfer function models with bounded noise. The contributions to the error bounds by noise, transient effects due to unmodeled disturbances, and modeling error due to model/system mismatch were all explicitly identified. Other work on developing algorithms for hard error bounds includes contributions by Giarre et al. (1997), Gunnarson (1993) and Zhu (1989), among others. Ackay and Ninness (1998)

quantified bounds in different norms for models developed using rational basis functions. Worst-cases system identification has been studied by Chen et al. (1995) and Dahleh et al. (1993). Regardless of the method, the hard bounding approaches result in error bounds which are highly conservative. They are necessarily so to ensure that even the worst case is enclosed by the bounds. Often, the bounds are unnecessarily conservative because conservative parameter space bounds can become even more conservative when transformed to transfer function space if the transformations are based on approximations to the true expectation surface defined by $f(\boldsymbol{x}, \boldsymbol{\theta})$.

Hard bounding approaches are often advocated on the basis that robust control theory requires strictly true and known error bounds. However, for real systems, there is always uncertainty about the model and the disturbances entering the system. Therefore, we believe it is unrealistic to propose methods to determine certain error bounds. To achieve near certainty, hard bounding methods overestimate the uncertainty and produce error bounds which are inappropriately wide. Goodwin et al. (1992) argued that hard error bounds are inappropriate because prior assumptions about noise, disturbances and control actions can never be known absolutely, and so the idea of certain limits is misguided. By this argument, a probabilistic approach is more consistent with the realities of system identification.

Soft error bounds are not guaranteed to contain the system performance, but rather are said to have a specified probability of containing the system performance. A soft approach to estimating error bounds for transfer function models was developed by Goodwin et al. (1992). They considered error due to bias from model/system mismatch and error due to noise in the measured data. The bias error was estimated by assuming a stochastic prior model for the distribution of the unmodeled dynamics. This model was embedded into the system estimation problem so that its parameters were estimated along with the parameters of the system model. These ideas were extended by Schoukens and Pintelon (1994) to allow for the case of colored noise in

continuous-time systems explored in the frequency domain. Like most work to date in this area, the method is restricted to models which are linear in the parameters. Goodwin et al. (1992) did apply their method to an ARMAX model by first linearizing the model. The authors claimed that their method performed well even for this case. The methods proposed in this paper can be used with a broad range of models, including nonlinear models.

DeVries and Van den Hof (1995) adopted a mixed deterministic/probabilistic approach to determining error bounds, accounting for three sources of uncertainty: undermodeling, noise disturbance and unknown initial conditions. The model error due to unmodeled dynamics and unknown past input signals were considered deterministic worst-case quantities, and the noise disturbance was considered to be stochastic. Their algorithm was a two-step approach whereby the bias was estimated first, and the noise uncertainty was estimated second. Again, only linear finite impulse response (FIR) models were considered.

Ninness and Goodwin (1995a) examined the relationship between bounded-error and stochastic estimation theory, and showed that for many problems, the two approaches are equivalent when a Baysian framework is used. A good overview of estimating uncertainty in models used for control is given by Ninness and Goodwin (1995b).

In this work, we adopt the belief that if a proposed model can not be rejected on the basis of graphical and numerical validation tests (Box and Jenkins, 1972; Ljung, 1987), then there is no basis for assuming a deficiency in the form of the model nor significant bias in the model estimates. The methods we propose account for parameter uncertainty only (except in the case of k-step-ahead predictions), implying that the form of the proposed model is sufficiently general to capture the dynamics of the true system.

7.4 The Likelihood Function and Estimation

If it is assumed that each random shock a_t is independently and identically normally distributed with zero mean and variance σ_a^2, then the likelihood function $L(\boldsymbol{\theta})$ for the unknown parameters $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}) = (2\pi\sigma_a^2)^{-n/2}|\boldsymbol{\Omega}|^{-1/2}\exp\left(\frac{-\boldsymbol{y}'\boldsymbol{\Omega}^{-1}\boldsymbol{y}}{2\sigma_a^2}\right) \tag{7.31}$$

where y_t is given by (7.1) and $\sigma_a^2\boldsymbol{\Omega}$ denotes the $n \times n$ covariance matrix for $\boldsymbol{y}$.

There are several possible approaches to estimating the parameters of a transfer function of the form given in (7.1). However, since generalized profiling is a likelihood-based approach to estimating uncertainty, all parameter estimation for the purposes of profiling has been done on the basis of maximum likelihood. In Section 7.5 we discuss profiling in the context of alternative estimation criteria (i.e. criteria other than maximum likelihood).

Several algorithms for computing the value of the likelihood function for transfer function models have been proposed. Lam and Watts (1991) based their profiling calculations on the expression for the exact likelihood function of an ARMA model develop by Ansley (1979). However, many others have proposed expressions and algorithms for computing the exact likelihood. Some of these include those of: Newbold (1974), Ali (1977), and Ljung and Box (1979). Harvey and Phillips (1979), and Åström (1980), among others, have developed algorithms based on the Kalman filter. Expressions for the exact likelihood function for vector ARMA(p,q) processes have been developed by: Osborn (1977), Phadke and Kedem (1978), Hillmer and Tiao (1979), and Nicholls and Hall (1979). These, of course, can also be used for the special case of univariate ARMA models.

We have considered four different algorithms to compute likelihood intervals. Two of the algorithms are exact maximum likelihood methods. One is based on the al-

gorithm proposed by Ansley (1979), and the other uses time-varying Kalman filtering (Åström,1980). The other two algorithms are pseudo-maximum likelihood approaches. The "approximate likelihood" method is based on the optimization of an approximate likelihood function with respect to the parameters *and* the unknown starting values. The "conditional likelihood" algorithm uses backcasting to compute unknown starting values (Box and Jenkins, 1976).

Ansley's algorithm is based on a clever transformation of an ARMA-type model. The procedure for using Ansley's algorithm in the context of transfer function models is given in Figure 7.2. However, we have found that full maximum likelihood estimation using Kalman filtering is more computationally efficient. Both methods produce the same estimates and are interchangeable. The results reported here have been computed using time-varying Kalman filtering. Only a very brief overview of Kalman filtering will be given here. For further details refer to Ljung (1987) and Åström (1980).

To use time-varying Kalman filtering, the transfer function must be expressed in state-space form. We consider a state-space model having the form:

$$\begin{aligned} \hat{x}(t+1) &= \Lambda(\boldsymbol{\theta})\hat{x}(t) + \Gamma(\boldsymbol{\theta})u(t) + w(t) \\ y(t) &= \Theta\hat{x}(t) + v(t) \end{aligned} \tag{7.38}$$

Let

$$\begin{aligned} E\{w(t)w^T(t)\} &= R_1(\boldsymbol{\theta}) \\ E\{v(t)v^T(t)\} &= R_2(\boldsymbol{\theta}) \\ E\{w(t)v^T(t)\} &= R_{12}(\boldsymbol{\theta}) \end{aligned} \tag{7.39}$$

1. Write the transfer function model in the form

$$y_t = \frac{B(q^{-1})}{A(q^{-1})F(q^{-1})}q^{-k}u_t + \frac{C(q^{-1})}{A(q^{-1})D(q^{-1})}a_t \tag{7.32}$$

2. Assume that all u_t are known, even for $t < 0$. Alternatively, assume that the process was operating at steady state prior to the experiment, i.e., $u_t = 0$, $t < 0$, where u_t is expressed in terms of deviations from set point.

3. Then, using initial estimates of the parameters, compute the filtered series

$$u_f = \frac{B(q^{-1})}{A(q^{-1})F(q^{-1})}q^{-k}u_t \tag{7.33}$$

4. Compute

$$w_t = y_t - u_f \tag{7.34}$$

5. Now write the model as a time series

$$\phi(q^{-1})w_t = \theta(q^{-1})a_t \tag{7.35}$$

where $\phi(q^{-1}) = A(q^{-1})D(q^{-1})$ and $\theta(q^{-1}) = C(q^{-1})$.

6. Ansley's transformation is

$$z_t = \begin{cases} w_t, & t = 1, \ldots, m \\ \phi(q^{-1})w_t, & t = m+1, \ldots, n \end{cases} \tag{7.36}$$

where $m = max(na \times nd, nc)$, and na, nb, nc and nd are the degrees of the polynomials $A(q^{-1})$, $B(q^{-1})$, $C(q^{-1})$ and $D(q^{-1})$, respectively. Let:

$$v_t = \theta(q^{-1})a_t = \phi(q^{-1})w_t \tag{7.37}$$

The series v_t is autocorrelated only up to lag q. Then, the covariance matrix for z_t has a maximum bandwidth of m for the first m rows and a bandwidth of nc thereafter (Ansley, 1979). See Ansley (1979) for an efficient way to compute the likelihood based on the transformed series z_t.

Figure 7.2: A step-by-step algorithm for maximum likelihood estimation of a SISO transfer function model (modified from Ansley, 1979).

Then, the time-varying Kalman filter is (Ljung, 1987):

$$\begin{aligned}
\hat{x}(t+1) &= \Lambda(\boldsymbol{\theta})\hat{x}(t) + \Gamma(\boldsymbol{\theta})u(t) + K(t)[y(t) - \Theta(\boldsymbol{\theta})\hat{x}(t)] \\
\hat{y}(t) &= \Theta\hat{x}(t) \\
K(t) &= [\Lambda(\boldsymbol{\theta})P(t)\Theta^T(\boldsymbol{\theta}) + R_{12}(\boldsymbol{\theta})][\Theta(\boldsymbol{\theta})P(t)\Theta^T(\boldsymbol{\theta}) + R_2(\boldsymbol{\theta})]^{-1} \\
P(t+1) &= \Lambda(\boldsymbol{\theta})P(t)\Lambda^T(\boldsymbol{\theta}) + R_1(\boldsymbol{\theta}) - K(t)[\Theta(\boldsymbol{\theta})P(t)\Theta^T(\boldsymbol{\theta}) + R_2(\boldsymbol{\theta})]K^T(t)
\end{aligned} \tag{7.40}$$

Because $K(t)$ and $P(t+1)$ are time varying, this formulation of the Kalman filter is able to deal with the transient properties of the estimation problem which are introduced when the state of the system prior to $t = 0$ is unknown. When implementing the Kalman filter, numerical problems can arise. There are several recommended approaches called factorization (Chui and Chen, 1987) methods which should be used in such cases. However, for the examples considered here, no numerical difficulties were encountered. In this work, Kalman filtering is used to compute an unconditional estimate of $\boldsymbol{e}$ which is then used to compute the value of the likelihood function. Åström (1980) gives a thorough discussion of the use of Kalman filtering to compute maximum likelihood estimates for parameters in a transfer function model.

The approximate likelihood algorithm is based on augmenting the number of parameters to be estimated. If the starting values for the series a_t and y_t were known, then the likelihood function could be written as:

$$L(\boldsymbol{\theta}) = (2\pi\sigma_a^2)^{-n/2} \exp\left(\frac{-\boldsymbol{e}^T\boldsymbol{e}}{2\sigma_a^2}\right) \tag{7.41}$$

where $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ is the $n \times 1$ vector of residuals. Then, the maximum likelihood estimates would be equal to the least squares estimates. We compute approximate MLEs by maximizing (7.41) with respect to the parameters and the initial conditions for a_t. The resulting estimates are not the true maximum likelihood estimates and they

can be severely biased when only a short data set is available (Thisted, 1988). The measures of nonlinearity introduced in Chapter 6 can be used to decide whether the vector of parameters is "close" to a stability/invertibility boundary. The conditional likelihood algorithm is also based on (7.41). In this algorithm, the initial conditions for a_t are estimated using backcasting (Box and Jenkins, 1976) and then the value of the likelihood function is computed using (7.41).

All calculation and visualization was done in MATLAB^{TM}. MATLAB^{TM} 's simplex algorithm was used to solve all unconstrained optimization problems, while MATLAB^{TM} 's sequential quadratic programming (SQP) algorithm was used to solve all constrained optimization problems.

7.5 Alternate Estimation Criteria

Profiling is based on the τ statistic, which in turn is based on the likelihood function for the parameters. Estimation based on the principle of maximum likelihood can be justified using statistical arguments; however, there are many other estimation criteria that can be justified by the intended end use of the model. For example, if a model is being developed for a process with dead time d, and this model is to be used to design a controller, then it may be appropriate to choose parameters which minimize the d-step-ahead predictions.

Here we attempt to identify some of the issues involved in using profiling in the context of models fitted using a multi-step-ahead prediction error criterion. The development of an explicit methodology for using profiling with other than maximum likelihood estimation is left for future work. When prediction errors are independently and identically normally distributed (for example, in the case of nonlinear regression models), the τ statistic can be written in terms of sums of squares of residuals as

follows:

$$\tau = sign(g(\boldsymbol{\theta}) - g(\hat{\boldsymbol{\theta}}))\sqrt{\frac{S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}})}{s^2}} \tag{7.42}$$

In this case, $S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}})$ can be shown to be independent of s^2 (Freund and Walpole, 1987), and consequently, the ratio $\frac{S(\boldsymbol{\theta})-S(\hat{\boldsymbol{\theta}})}{s^2}$ can be shown to follow an F distribution with one and $n-p$ degrees of freedom.

When estimating parameters on the basis of k-step-ahead predictions, we minimize

$$\bar{S}(\boldsymbol{\theta}) = \sum_{t=1}^{n}(y_t - \hat{y}_{t|t-k})^2 \tag{7.43}$$

where $\hat{y}_{t|t-k}$ is the optimal k-step-ahead prediction of y_t given information up to and including time $t-k$. One approach to using profiling in this context would be to compute

$$\bar{\tau} = sign(g(\boldsymbol{\theta}) - g(\hat{\boldsymbol{\theta}}_{alt}))\sqrt{\frac{\bar{S}(\boldsymbol{\theta}) - \bar{S}(\hat{\boldsymbol{\theta}}_{alt})}{s^2}} \tag{7.44}$$

where $\hat{\boldsymbol{\theta}}_{alt}$ is the vector of parameter values which minimizes $\bar{S}(\boldsymbol{\theta})$. As above, s^2 is an estimate of the variance of the white noise process. However, when using an alternate estimation criterion, it may be appropriate to use an alternate estimate for s^2, for example $s^2 = \bar{S}(\boldsymbol{\theta})/(n-p)$. To use $\bar{\tau}$ to compute uncertainty intervals would require that its sampling distribution be known. The task of deriving an expression for the distribution of $\bar{\tau}$ is complicated by the fact that the residuals $\bar{e}_t = y_t - \hat{y}_{t|t-k}$ are autocorrelated, and $\bar{S}(\boldsymbol{\theta}) - \bar{S}(\hat{\boldsymbol{\theta}}_{alt})$ and s^2 are not necessarily independent.

The key to deriving the distribution of $\bar{\tau}$ may lie in the fact that to minimize $\bar{S}(\boldsymbol{\theta})$ is to solve the weighted least squares problem

$$\text{Minimize } \boldsymbol{eWe} \tag{7.45}$$

where $\boldsymbol{W}$ is a square matrix of weights. The weights may also be functions of the parameters. Mathai and Provost (1992) discussed the distribution of ratios of quadratic forms in random variables. It is by exploiting the properties of quadratic forms that we foresee the use of profiling in the context of alternate estimation criteria.

7.6 Illustrative Examples

All examples considered in this section are based on a second-order discrete system which is described by the model

$$y_t = \frac{0.0172}{1 - 1.7236q^{-1} + 0.7408q^{-2}} q^{-2} u_t + \frac{1}{(1 - 0.90q^{-1})} a_t \tag{7.46}$$

We will refer to (7.46) as Model 1. As noted by Shirt et al. (1994), the parameters of Model 1 are close to the stability boundaries for a second-order system. In PACF space (see Chapter 6) the vector of parameters is $\boldsymbol{\phi}^T = [0.0172,\ 0.90,\ 0.9901,\ -0.7408]$. With this parameterization it is easy to see that the vector of parameters is close to one side of the stability/invertibility region which is a hypercube with sides at ± 1. Lam and Watts (1989) showed that, for ARMA models, parameters near the stability/invertibility boundaries tend to show significant nonlinearity. One purpose of the current study is to identify whether linearization confidence intervals are reliable for control-relevant statistics for this system. It is useful, then, to consider a system for which they might reasonably be expected to be poor. The variance of the white noise input σ_a^2 was chosen to be equal to 0.0361. This results in the disturbance $\frac{1}{1-0.9q^{-1}} a_t$ having a variance of 0.19. The total variance of the simulated y_t series was 0.51. The system, then, has a high signal to noise ratio. We will show that despite this high ratio, the parametric uncertainty and nonlinearity are significant for the several of the functions of parameters we consider.

In examples where a controller is considered, we employ one of two controllers.

Controller 1 is a Dahlin controller (Seborg, 1989) designed such that, in the absence of disturbances, the closed loop dynamics of the system are expected to follow a first-order-plus-dead-time model. This controller has the form

$$G_c(q^{-1}) = \frac{(1-\gamma)q^{-d}}{1-\gamma q^{-1}-(1-\gamma)q^{-d}} \frac{1}{G_p^*(q^{-1})} \tag{7.47}$$

where γ is a tuning constant which is related to the discrete time constant of the desired closed loop system, and

$$G_p^*(q^{-1}) = \frac{B^*(q^{-1})}{F^*(q^{-1})}q^{-d} = \frac{0.0172}{1-1.7236q^{-1}+0.7408q^{-2}}q^{-2} \tag{7.48}$$

is the true process model. For our example we choose $\gamma = 0.5$. Although selecting $\gamma = 0.5$ results in a very aggressive controller, the controller is useful for illustrative purposes. In practice, values of γ in the range 0.95-0.98 would be more reasonable since they would result in a controller which would call for moderate changes in the manipulated variable. With $\gamma = 0.5$

$$G_c(q^{-1}) = \frac{0.5F^*(q^{-1})}{1-0.5q^{-1}-0.5q^{-2}B^*(q^{-1})} \tag{7.49}$$

$$= \frac{29.07(1-1.7236q^{-1}+0.7408q^{-2})}{1-0.5q^{-1}-0.5q^{-2}} \tag{7.50}$$

Note that the controller has a pole at $q^{-1} = 1$, causing it to have integral action.

In practice, $G_p^*(q^{-1})$is never known exactly, but rather is estimated by $\hat{G}_p(q^{-1})$, where $\hat{G}_p(q^{-1})$ is assumed to have a form such that it can capture the true process dynamics, even though the parameter values are only estimates of the true parameters values. Therefore, the controller is implemented as

$$\hat{G}_c(q^{-1}) = \frac{0.5\hat{F}(q^{-1})}{(1-0.5q^{-1}-0.5q^{-2})\hat{B}(q^{-1})} \tag{7.51}$$

In the section on prediction (Section 7.6.4) we consider a multi-step-ahead control algorithm (Ydstie et al., 1985) and we refer to it as Controller 2.

A generalized binary noise (GBN) test signal (Tulleken, 1990) was used for illustrative purposes. GBN test signals are sequences of inputs alternating between two levels, where the switching of the signal from the low level to the high level or vice versa is governed by a probability $p_{switching}$. The parameter $p_{switching}$ is the probability at any sampling point that the signal will remain at the same level. The value of $p_{switching}$ was chosen to 0.9, and the levels switched between plus and minus one. The length of the data set used for identification can have a profound affect on the quality of the estimates of parameters and functions of parameters, as can the spectrum of the input sequence. For the purposes of identification and inference, we used a simulated data set consisting of 500 observations of the process.

7.6.1 The Parameters

The first example of a function of parameters we will consider is the individual parameter f_1 in the model

$$y_t = \frac{b_0}{1 + f_1 q^{-1} + f_2 q^{-2}} u_{t-d} + \frac{1}{1 + d_1 q^{-1}} a_t \tag{7.52}$$

That is, we choose

$$g(\boldsymbol{\theta}) = f_1 \tag{7.53}$$

The profile t plot for f_1 is shown in Figure 7.3. See Table 7.1 for the estimation results for all of the parameters of the fitted model.

Although a reasonably long data set was used for identification, the profile t plot is significantly nonlinear and the 95% linearization confidence interval [-1.92 -1.44] differs from the 95% likelihood interval [-1.85 -1.31]. The difference between the

two intervals becomes more pronounced as the level of confidence increases (i.e. as α decreases). To rely on the linearization approximations for this example would be to underestimate the uncertainty in the estimate of f_1 with regard to values greater than the maximum likelihood estimate, and to overestimate the uncertainty for values of f_1 less than the MLE. For this example ζ_{min} (see Chapter 6) was equal to 1.5, indicating moderate nonlinearity. The parameter f_2 was also nonlinear, having a 95% linearization interval [0.48 0.91] which differed from the 95% likelihood interval [0.34 0.87]. As stated above, the signal to noise ratio was relatively high in this example. The lower the ratio, the more uncertainty will be present in the estimate of a function of parameters. This uncertainty translates into wider likelihood intervals. Often, the nonlinearity increases as the signal to noise ratio decreases, although this is not necessarily so, and we advocate the use of profiling to ensure that significant nonlinearity is accounted for in any case.

Table 7.1: Table of Maximum Likelihood Estimation Results.

Model	Input	s^2	Parameter	True Value	MLE
1	1	0.036	b_1	0.0172	0.0199
			d_1	-0.9	-0.8878
			f_1	-1.7236	-1.6783
			f_2	0.7408	0.6989

In the statistical literature the emphasis has been on quantifying uncertainty in estimates of parameters of models, and on designing experiments to improve those estimates. Control specialists have long recognized that this emphasis is not always consistent with their needs. Most often in control applications, the values of the parameters themselves are of little interest. Usually the objective is to develop a model which captures some aspect of the behavior of the true system. A flexible control-relevant approach to designing experiments was presented in Ljung (1987).

In the work for this paper, the emphasis is on illustrating the use of generalized profiling for control-relevant functions of parameters. In control, decisions are based on functions of parameters and not on the parameters themselves; therefore it is more important to identify reliable likelihood intervals for the functions of parameters of interest rather than for the parameters themselves. This is especially true since the uncertainty and nonlinearity of individual parameters is not necessarily reflected in a function of those parameters.

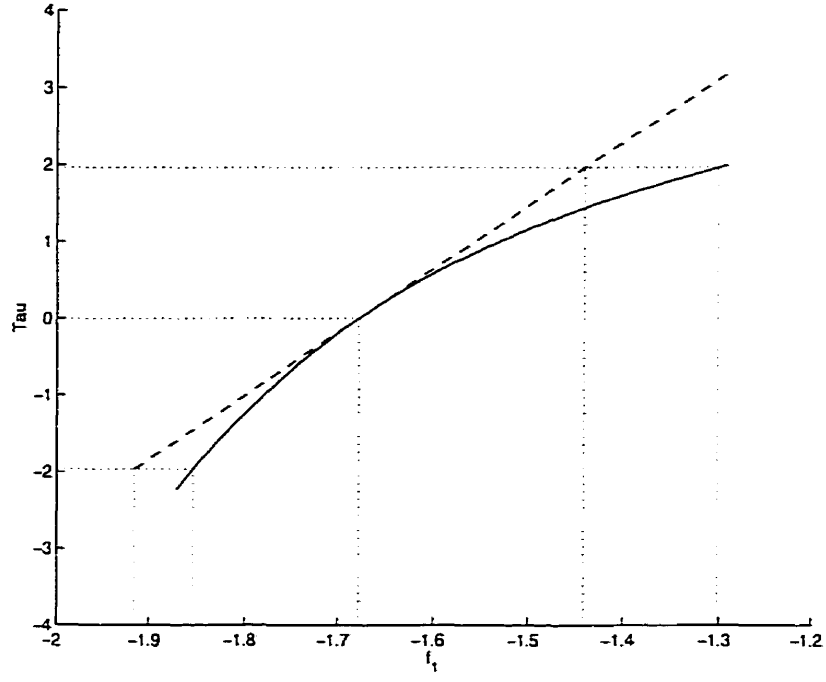

Figure 7.3: Profile t plot for parameter f_1.

7.6.2 Steady-State Gain

All three of the parameters in the process model (7.52) showed some nonlinearity. However, it does not follow that all functions of these parameters will also be nonlinear. The value of the steady-state gain of the process described by this model is

given by the expression

$$gain = \frac{b_0}{1 + f_1 + f_2} \tag{7.54}$$

Therefore, the gain is a function of the model parameters. The profile t plot for the gain of this model is shown in Figure 7.4. Although the individual parameters are nonlinear, the gain behaves relatively linearly and the linearization confidence interval for this function of parameters is a good approximation of the corresponding likelihood interval. This suggests that the nonlinearity of this estimation and inference problem could be reduced by choosing to fit a model of the form:

$$y_t = \frac{gain(1 + f_1 + f_2)}{1 + f_1 q^{-1} + f_2 q^{-2}} u_{t-d} + \frac{1}{1 + d_1 q^{-1}} a_t \tag{7.55}$$

instead of the model given in (7.52). It is not uncommon for a function of parameters to behave linearly when one or more of the individual parameters behave nonlinearly, or conversely, for a function of parameters to behave nonlinearly when all of the parameters behave linearly (Quinn et al., 1999a; Chapter 3). We hope that this work highlights the need to consider each model and function of parameters individually, and that it emphasizes the need to exercise caution when making assumptions about the appropriateness of linearization confidence intervals.

7.6.3 Gain Margin

The gain margin was defined in Section 7.2. It is an important control-relevant statistic used to measure robustness and stability (Palmor and Shinnar, 1981). Typically, Bode plots are constructed for the plant-plus-controller system as operated in open loop. That is, the transfer function

$$G_{OL}(q^{-1}) = G_c(q^{-1}) G_p(q^{-1}) \tag{7.56}$$

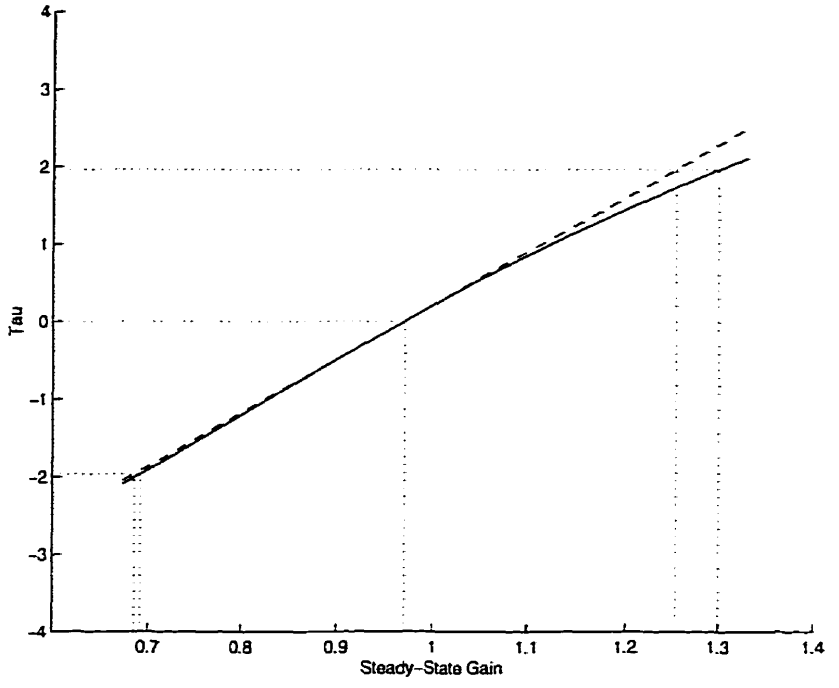

Figure 7.4: Profile t plot for the steady-state gain.

is of primary interest, where $G_c(q^{-1})$ and G_p are defined in (7.49) and (7.48), respectively.

For this example we consider model (7.52) and Controller 1, so that

$$\hat{G}_{OL}(q^{-1}) = \frac{0.5\hat{F}(q^{-1})}{(1-0.5q^{-1}-0.5q^{-2})\hat{B}(q^{-1})}\hat{G}_p(q^{-1}) \tag{7.57}$$

$$= \frac{0.5}{1-0.5q^{-1}-0.5q^{-2}} \tag{7.58}$$

In practice, a controller is designed based on the "best" process model $\hat{G}_p(q^{-1})$. Once designed and implemented, we assume that $\hat{G}_c(q^{-1})$ is fixed (i.e., the controller is not adaptive). $\hat{G}_{OL}(q^{-1})$ is computed based on the maximum likelihood estimates of the model parameters; therefore, when the controller is implemented, we expect that $G_{OL}(q^{-1})$ may not behave exactly as planned because $\hat{G}_p(q^{-1})$ is only an estimate of the true process (Palmor and Shinnar, 1979).

To profile the gain margin of $G_{OL}(q^{-1})$ we consider the controller to be fixed and examine only the influence of the uncertainty in $G_p(q^{-1})$ on the gain margin.

We are interested in knowing what values of the gain margin are plausible once a given controller with fixed parameter values is implemented on the real process. Our emphasis is on the fact that we know exactly what controller is to be implemented on the process, but we don't know the process exactly. Although we assume the process to be "fixed" in so far as we assume that the true system does not change over time, we do not have an exact model for the process. It is the uncertainty in the process model that contributes to the uncertainty in the estimate of the gain margin.

A profile t plot of the gain margin of $G_{OL}(q^{-1})$ is shown in Figure 7.5. For this example, the 95% linearization confidence interval [0.88 5.12] includes values less than one, which indicates that at the 95% confidence level, there is statistical evidence to suggest that the closed loop process may be unstable. However, the lower limit of the 95% likelihood interval is 1.37, leading to the conclusion that the closed-loop system will be stable. The upper limit of this interval is 6.64. The controller plus the true process is stable with a gain margin of 3.50. Especially when safety is a critical issue, it is important to have reliable estimates of the uncertainties of any statistics of interest.

So far we have considered one case, the case in which the parameters of the controller are fixed, and uncertainty in the observed data propagates to the estimate of the gain margin through the parameters of the process model. There are two other cases that may be of interest in some applications.

In Case 2, the values of the parameters in the process model are "fixed", and the uncertainty in the gain margin arises from uncertainty in the values of the parameters in the controller. In Case 3, the values of the parameters in both the model of the process and the model of the controller are considered to be uncertain.

The inference results obtained in Case 2 will, in general, not be the same as those that would be obtained by fixing the values of the parameters in the controller but considering the parameters of the process model to be uncertain (Case 1). To

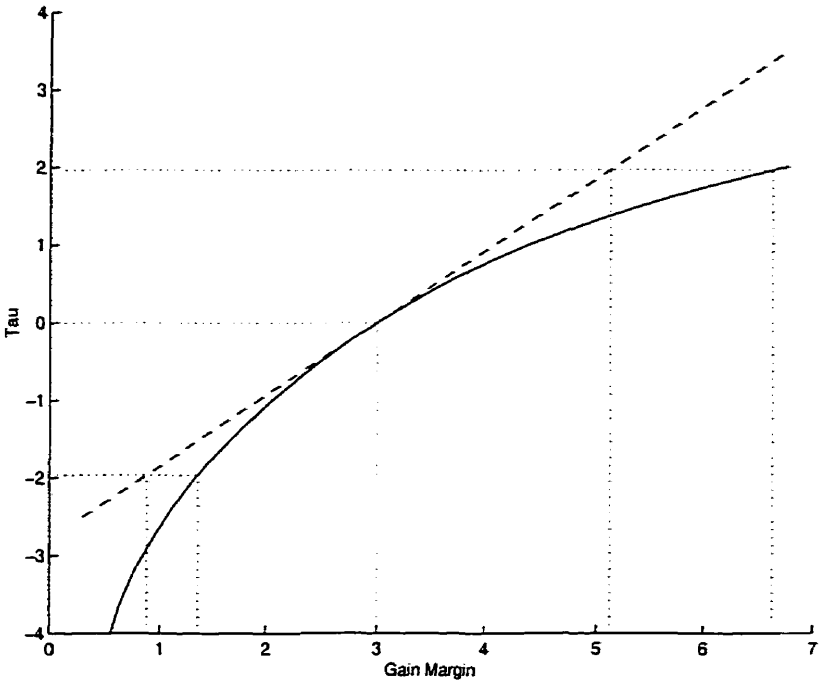

Figure 7.5: Profile t plot for the gain margin of the Model 1 based a Dahlin controller.

illustrate that different results can be obtained, we constructed a profile t plot for Case 2. The 95% likelihood interval for $G_{OL}(q^{-1})$ for this case was [1.32 5.94], which is quite different from the interval found for Case 1.

In some cases, it may be of interest at the design stage to consider the full range of plausible values of the gain margin in light of uncertainties in both the parameters in the controller transfer function and the parameters in the model (i.e. to consider Case 3). That is, we may be interested in the question: given a set of plausible controllers and a set of plausible models, what is the range of plausible values for the gain margin?

Conceptually, we are imagining the following scenario. If we were to perform a sequence of identification experiments, we would identify a set of estimates for the parameters from each experiment. From each set of parameter estimates $\boldsymbol{\theta}_1$ we would design a controller with parameter values $\boldsymbol{\theta}_2$. In this way we would identify a set of plausible values for the parameters of the process model and a corresponding set of plausible values for the parameters in the controller. In the case of the Dahlin

controller, the two sets are the same since the parameters that enter the process model are the same as those that enter the controller transfer function. These ideas are illustrated in Figure 7.6. Here we show that if we identify the set of values represented by point A in the set of process parameters, then we would design a controller having values represented by point A' in the set of controller parameters.

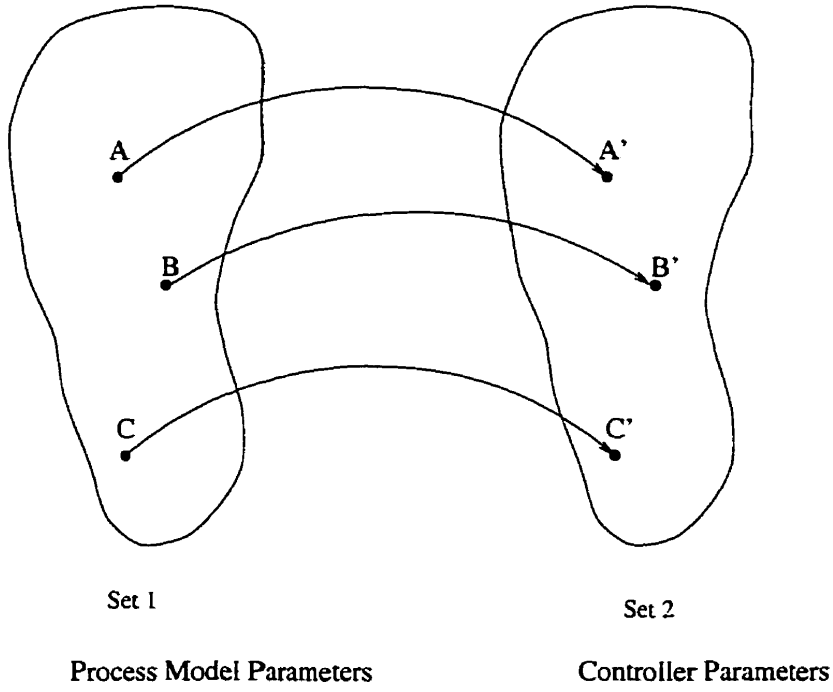

Figure 7.6: An illustration of the mapping of values in the set of process parameters to resulting values in the set of controller parameters.

It is possible, however, that a controller having parameter values represented by point A' would actually be implemented on a true process process that has parameters with values actually represented by point C. We are interested in knowing the range of possible values values for the gain margin, given that any combination of process parameters and controller parameters from within their respective allowable sets is possible. This is a fundamentally different problem than Case 1 where the values of the controller parameters were fixed.

For the problem involving a set of controller transfer functions and a set of models, profiling in its standard form can not be used. Difficulties arise from the constraint

imposed by the controller design algorithm. If we attempt to profile the gain margin using the standard algorithm, then we find that the value of the gain margin will always be equal to its maximum likelihood estimate. This is because no matter what values of the parameters we may consider for the process, we will design a controller to produce the desired closed-loop behavior, the behavior that is characterized by the maximum likelihood estimate of the gain margin.

Case 3 is a combinatorial problem in which we are looking for all possible values of the gain margin arising from all possible combinations of allowable values of the process parameters and controller parameters. Such problems are better solved using resampling methods.

7.6.4 Prediction

In all areas of chemical engineering, models are fitted to data and subsequently used to make predictions . When dynamic models are being used, several scenarios can arise:

1. A pure time series model or transfer function model is identified. The data used in the identification are also used in making the predictions.

2. As in Scenario 1, a dynamic model is identified, but in this case, the prediction is made using a set of input/output data other than the one used to fit the model.

3. An input/output model is developed for the purposes of control. In this case, there may be several additional computational steps required before making a prediction, and it is possible that the form of the disturbance model will be changed to design the controller and make predictions. Control problems can be further sub-classified as:

(a) regulation problems, whereby the objective is to reject disturbances entering the system

(b) tracking problems, whereby the objective is to follow changes in the set-point.

Especially in model-based control, the accuracy and precision of the model predictions have direct influence on controller performance. Therefore, it is important to know the uncertainty in predictions. Ideally, this information should be available at the design stage. Profiling may be used to estimate uncertainty in k-step-ahead predictions made in any of the cases outlined above; however, it is important to first clearly define the problem and decide which sources of uncertainty are important.

There are several ways to formulate an expression for a k-step-ahead prediction (Åström and Wittenmark, 1990). Using the Diophantine identity, an expression for the k-step-ahead prediction can be developed as follows. Rearrange the general transfer function equation given in (7.1) so that it is written as (Ljung, 1987)

$$L(q^{-1})y_t = M(q^{-1})u_{t-d} + N(q^{-1})a_t \tag{7.59}$$

where

$$L(q^{-1}) = A(q^{-1})F(q^{-1})D(q^{-1}) \tag{7.60}$$

$$M(q^{-1}) = B(q^{-1})D(q^{-1}) \tag{7.61}$$

and

$$N(q^{-1}) = F(q^{-1})C(q^{-1}) \tag{7.62}$$

Then, use the Diophantine identity

$$N(q^{-1}) = L(q^{-1})P(q^{-1}) + q^{-k}Q(q^{-1}) \tag{7.63}$$

where $P(q^{-1})$ is a polynomial of degree $k-1$, to decompose $N(q^{-1})$ into past, present and future terms (Åström and Wittenmark, 1990). The k-step-ahead prediction is

$$\hat{y}_{t+k|t} = \frac{P(q^{-1})M(q^{-1})u_{t-d+k} + Q(q^{-1})y_t}{N(q^{-1})} \tag{7.64}$$

For $t-d+k < 0$ assume $u_{t-d+k} = 0$. In other words, since u_t is a deviation variable, we are assuming that the input to the system was constant prior to $t = 1$. For $k-d > 0$, use the values of u_{t+k-d} chosen by the multi-step-ahead model predictive controller described later in this section.

The "prediction error" is

$$e_{t+k|t} = P(q^{-1})a_{t+k} \tag{7.65}$$

which is a weighted sum of the sequence of future random errors. The expression in (7.65) is based on the assumption that the parameters are known perfectly. In other words, uncertainty in the parameters does not contribute to the uncertainty in the prediction.

We want to use profiling to estimate likelihood intervals for $\hat{y}_{t+k|t}$ that will take into account parameter uncertainty. Prior work in this area (Reinsel, 1980; Reimers, 1995) has been limited to models which are linear in the parameters. We propose the use of profiling to account for parameter uncertainty when the model of interest is nonlinear in the parameters. To follow the profiling procedure given in Figure 7.1, and to compute the profiling statistic τ in the usual way (i.e., using Equation 7.23), is to compute likelihood intervals for the *mean* or expected value of the k-step-ahead

prediction. By this approach we are accounting for parameter uncertainty in the prediction but we are neglecting the uncertainty due to the unknown future random errors. That is, we are neglecting $P(q^{-1})a_{t+k}$.

To establish a means by which to account for both the uncertainty due to parameter estimation and the uncertainty due to unknown values of future random shocks, we look to the theory on inference about predictions from linear regression models. For linear regression models, a $(1-\alpha)100\%$ confidence interval for the prediction of the *mean* value of the model is

$$\hat{y} \pm \; s\, t(n-p;\alpha/2)\sqrt{\boldsymbol{x}_0^T(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{x}_0} \tag{7.66}$$

where $\boldsymbol{x}_0$ is a vector of values for the regressor variables determined by the conditions for which the prediction is required. This expression does not account for the unknown future random shock a_{t+1}, just as standard profiling of $\hat{y}_{t+k|t}$ would not account for the future error $P(q^{-1})a_{t+k}$. However, a $(1-\alpha)100\%$ confidence interval for a specific future value of a steady-state linear system is given by

$$\hat{y} \pm t(n-p,\alpha/2)\sqrt{s^2 + s^2\boldsymbol{x}_0^T(\boldsymbol{V}^T\boldsymbol{V})^{-1}\boldsymbol{x}_0} \tag{7.67}$$

This expression accounts for the uncertainty due to the unknown future random shock a_{t+1} by adding to the square of the standard error of $\hat{y}$ the term s^2 which is the variance of a_{t+1}. When the model and the function of parameters are both linear in the parameters

$$\left|\frac{g(\boldsymbol{\theta}) - g(\hat{\boldsymbol{\theta}})}{s\sqrt{\frac{dg}{d\boldsymbol{\theta}}^T\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}(\boldsymbol{V}^T\boldsymbol{V})^{-1}\frac{dg}{d\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}}}\right| = \sqrt{\frac{S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}})}{s^2}} \tag{7.68}$$

and

$$s^2 \left.\frac{dg}{d\boldsymbol{\theta}}^T\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{V}^T\boldsymbol{V})^{-1} \left.\frac{dg}{d\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \frac{s^2(g(\boldsymbol{\theta}) - g(\hat{\boldsymbol{\theta}}))^2}{S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}})} \tag{7.69}$$

Then, add the expression for the variance of the future random error to both sides so that

$$s^2 \left(\frac{\sigma_p^2}{s^2} + \left.\frac{dg}{d\boldsymbol{\theta}}^T\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{V}^T\boldsymbol{V})^{-1} \left.\frac{dg}{d\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right) = s^2 \left(\frac{\sigma_p^2}{s^2} + \frac{(g(\boldsymbol{\theta}) - g(\hat{\boldsymbol{\theta}}))^2}{S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}})}\right) \tag{7.70}$$

where σ_p^2 is the variance of the term accounting for the contribution of the unknown future errors to the prediction. For regression models, $\sigma_p^2 \approx s^2$, and for transfer function models

$$\sigma_p^2 = var(e_{t+k|t}) \approx s^2(1 + \sum_{i=1}^{k-1} p_i^2) \tag{7.71}$$

where p_i represents the i^{th} coefficient of the polynomial $P(q^{-1})$. Therefore,

$$\tau_{pred} = sign(g - \hat{g}) \sqrt{\frac{S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}})}{\sigma_p^2 \frac{(S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}}))}{(g(\boldsymbol{\theta}) - g(\hat{\boldsymbol{\theta}}))^2} + s^2}} \tag{7.72}$$

Likelihood intervals based on τ_{pred} do account for both parameter uncertainty and uncertainty due to unknown future errors. When $\sigma_p^2 = 0$ (i.e., the variance of the error due to unknown future noise is zero), then (7.72) reduces to the expression for τ given in (7.23). Both τ and τ_{pred} can be appropriate for use when profiling multi-step-ahead predictions, depending on the sources of error that are of interest. For example, when the disturbances are expected to have a significant affect on the process, then τ_{pred} would be appropriate. Conversely, in the case where we are using control to track setpoint changes, it may be reasonable to assume that the effects of the disturbances are negligible relative to the magnitudes of the effects of setpoint changes. Then, we

would not be interested in uncertainty due to future random errors and would only be concerned with parametric uncertainty. In this case it would be appropriate to use the standard profiling algorithm based on τ. The standard profiling algorithm is also appropriate for use any time one is interested in the expected (mean) value of a multi-step-ahead prediction.

To illustrate the use of profiling to compute likelihood intervals for predictions, we consider an example of a prediction made in the context of regulatory control. The other scenarios described at the beginning of this section are all subsets or variations of this problem. Especially in this case, it is important to outline the problem and the assumptions being made, because different disturbance models are often used at different stages of the identification-design-prediction problem, and the variance of the white noise inputs may also change from stage to stage. A step-by-step procedure for our example is given in Figure 7.7.

We considered a system described by Model 1. Using this model and the GBN input sequence, 500 observations of the system were generated. The input and output data are shown in Figure 7.8. The simulated data were used to estimate the parameters of a model having the same form as the "true" model. The results of the full maximum likelihood identification were given earlier in Table 7.1.

The next step in this case study was the design of a controller. For the purposes of designing the controller we assumed that the disturbances would follow a random walk model. That is, we use the transfer function

$$y_t = \frac{0.0199}{1 - 1.6783q^{-1} + 0.6989q^{-2}} q^{-2} u_t + \frac{1}{(1 - q^{-1})} a_t \tag{7.79}$$

to design the controller. The fitted disturbance model was replaced by a random walk model to ensure that the controller would have integral action, and consequently, zero offset.

1. Generate a GBN input sequence of length 500, alternating between values of ± 1.

2. Generate an iid N(0, 0.036) white noise sequence of length 500.

3. Using the data generated in Steps 1 and 2, and the model

$$y_t = \frac{0.0172}{1 - 1.7236q^{-1} + 0.7408q^{-2}} q^{-2} u_t + \frac{1}{(1 - 0.90q^{-1})} a_t \tag{7.73}$$

simulate 500 values of y_t. Equation 7.73 is referred to as the "true" model.

4. Using the simulated data, estimate the parameters of a model having the form:

$$y_t = \frac{b_0}{1 + f_1 q^{-1} + f_2 q^{-2}} u_{t-2} + \frac{1}{1 + d_1 q^{-1}} a_t \tag{7.74}$$

Let

$$\hat{G}_P(q^{-1}) = \frac{\hat{b}_0}{1 + \hat{f}_1 q^{-1} + \hat{f}_2 q^{-2}} \tag{7.75}$$

be the "fitted" process model, and

$$\hat{G}_D(q^{-1}) = \frac{1}{1 + \hat{d}_1 q^{-1}} \tag{7.76}$$

be the "fitted" disturbance model.

5. Using the fitted process model and a random walk model for the disturbance, design a controller using the algorithm described in Ydstie et al. (1985).

6. Generate an iid N(0, 0.0038) white noise sequence of length 100 called a_{closed}.

7. Using the model

$$\begin{aligned} y_t &= G_P(q^{-1}) u_{t-2} + \frac{1}{(1 - 0.99q^{-1})} a_t \\ &= \frac{0.0172}{1 - 1.7236q^{-1} + 0.7408q^{-2}} q^{-2} u_t + \frac{1}{(1 - 0.99q^{-1})} a_t \end{aligned} \tag{7.77}$$

the white noise sequence generated in Step 6, and the controller designed in Step 5, simulate the closed-loop system for 100 sampling intervals. Call the output series y_{closed} and the input sequence u_{closed}.

8. Using u_{closed}, y_{closed}, and the model

$$y_t = \hat{G}_P(q^{-1}) + \frac{1}{(1 - 0.99q^{-1})} a_t \tag{7.78}$$

make a k-step-ahead prediction.

9. Use the standard profiling algorithm to compute likelihood intervals for the k-step-ahead prediction. The data set used to estimate the parameters, and a model of the form given in (7.74), were used to compute the values of the likelihood function during profiling.

10. Use a profiling algorithm based on τ_{pred} to compute a likelihood interval for a new k-step-ahead prediction. Note that when computing σ_p^2, the variance of the white noise refers to the variance of the white noise used to generate the data being used for prediction. In this case $s_a^2 = 0.0038$.

Figure 7.7: The step-by-step procedure for simulating the closed loop system and measuring uncertainty in k-step-ahead predictions.

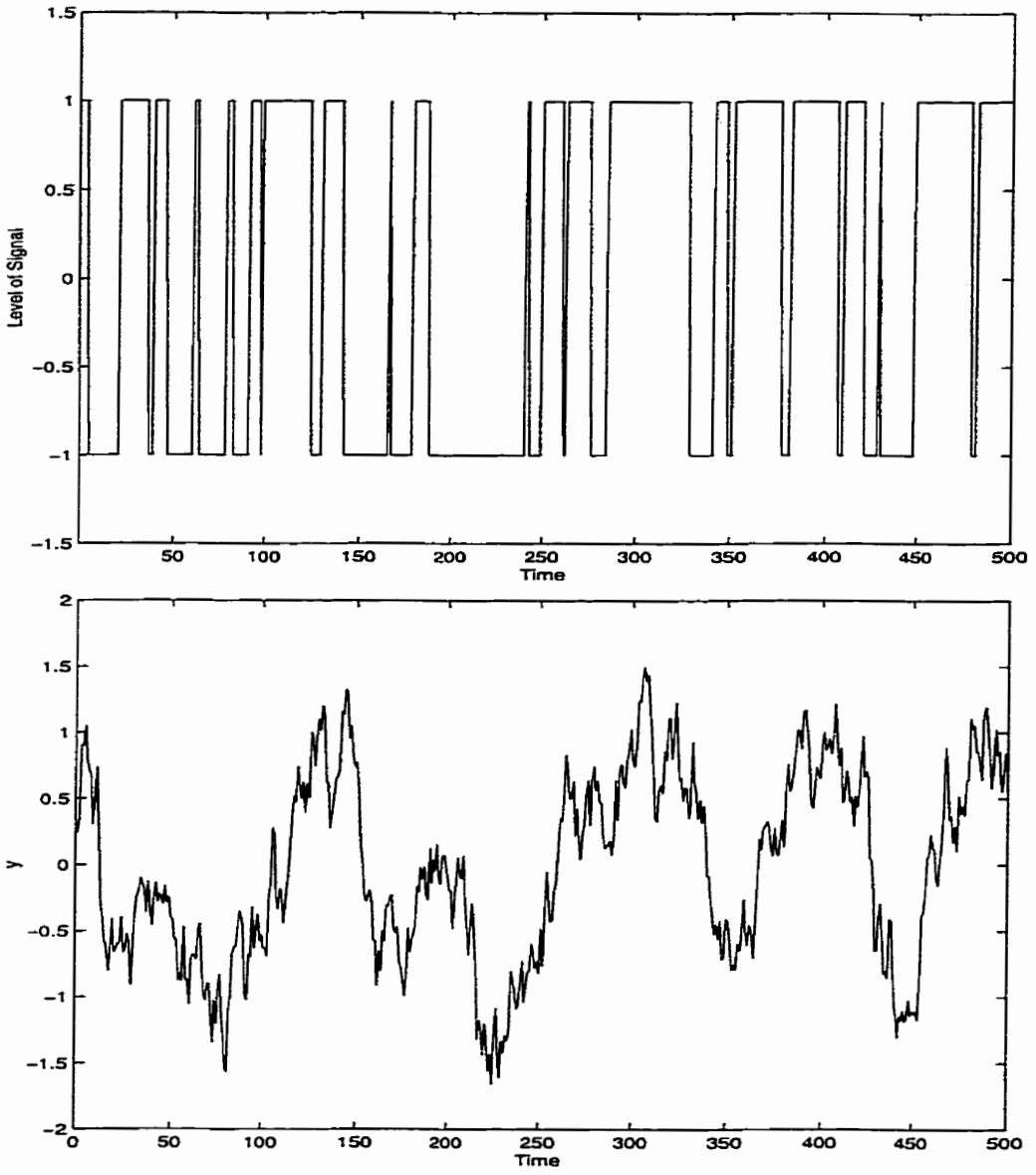

Figure 7.8: Simulated data used for identification purposes.

There are many ways to design a model predictive controller (Seborg et al., 1989). For the purposes of illustrating the profiling methodology, we consider one of the controllers proposed by Ydstie et al. (1985). This multi-step-ahead model predictive controller chooses the constant control action $u = u_t = \ldots = u_{t+\ell-d}$ such that the ℓ-step-ahead prediction will be equal to y^{target}.

Using the expression for the prediction given in (7.64) and the model given in (7.79), the input u which results in $\hat{y}_{t+\ell} = y^{target}$, and which minimizes the variance of the inputs, is

$$u = u_t = \frac{(y_t^{target} - h_{t+\ell})}{\sum_{j=1}^{\ell-d+1} \beta_j} \tag{7.80}$$

where

$$h_{t+\ell} = \sum_{i=0}^{n-1} \alpha_i y_{t-i} + \sum_{i=1}^{n-1} \beta_{\ell-d+i} u_{t-i} \tag{7.81}$$

$\alpha_0 = 1$, α_i is the i^{th} impulse response coefficient of $\frac{Q(q^{-1})}{N(q^{-1})}$ and β_i is the i^{th} impulse response coefficient of $\frac{P(q^{-1})M(q^{-1})}{N(q^{-1})}$(Ydstie et al., 1985). Note that $h_{t+\ell} = \hat{y}_{t+\ell|t}$ with $u_t = u_{t+1} = \ldots = u_{t+\ell-d} = 0$. In all cases, we choose $y^{target} = 0$ (i.e. we are considering regulation, not setpoint tracking). The prediction horizon ℓ changes from example to example. Once we compute $u_t = u_{t+1} = \ldots = u_{t+\ell-d}$ for control purposes, we use these same values in the calculation and profiling of multi-step-ahead predictions.

Once the parameters of the model were identified and the controller designed, we turned our attention to making predictions and measuring the uncertainty in those predictions. We consider that the system is operating in closed-loop, and at some point in the future we wish to make a multi-step-ahead prediction and estimate its uncertainty. The closed-loop system was simulated for a period of 100 sampling intervals. For this simulation we employed the true process model but we continued to assume that the disturbances followed a random walk model. However, we chose

to use a disturbance model which was close to being a random walk but which was stable so that its statistical properties would be known. We based our simulation on the transfer function

$$y_t = \frac{0.0172}{1 - 1.7236q^{-1} + 0.7408q^{-2}} q^{-2} u_t + \frac{1}{(1 - 0.99q^{-1})} a_t \tag{7.82}$$

Note that the variance of the white noise process $\{a_t\}$ used to generate the data used for the identification was chosen to be 0.0361 so that the variance of the disturbance process would be approximately 0.19. Because the disturbance model changed over the course of the case study, the variance of the white noise process was also changed as appropriate. For the closed-loop simulation, we employed an $\{a_t\}$ sequence having a variance of 0.0038 so that the variance of the disturbance described by

$$\frac{1}{1 - 0.99q^{-1}} a_t \tag{7.83}$$

would have a variance of 0.19,the same as the variance of the disturbance used in the identification step.

Given the information from the closed-loop system up to and including $t = 100$, we made several multi-step-ahead predictions. Although the closed-loop system was simulated based on (7.82), the true description of the process, in practice, would not be known and the prediction would be made based on the fitted process model. For the purposes of computing multi-step-ahead predictions for our closed-loop system we employed the system model used to design the controller which is given in (7.79).

The profiling of multi-step-ahead predictions can be based on τ or τ_{pred}, depending on which sources of uncertainty are of interest. We have computed likelihood intervals based on both. When profiling, the likelihood function is computed based on the data used to estimate the parameters, and a model of the form given in (7.52). When using

a profiling algorithm based on τ_{pred}, the value of σ_p^2 must be computed. We used the expression

$$\sigma_p^2 = var(e_{t+k|t}) \approx s^2(1 + \sum_{i=1}^{k-1} p_i^2) \tag{7.84}$$

which was developed earlier from (7.65) and (7.71). Care should be taken when computing s^2. In the context of σ_p^2, s^2 represents an estimate of the variance of the white noise process which generated the data being used to make the prediction. In our example, we used the model in (7.82), together with the closed-loop data, to compute a vector of residuals, and subsequently to estimate the variance. We used $s^2 = 0.0039$ to compute σ_p^2, which is different than the variance of the white noise used to generate the data for the identification step.

The case study was carried out twice for two different control horizons: $\ell = 4$ and $\ell = 8$. The closed-loop input and output data for the two cases are shown in Figures 7.9 and 7.10.

For the controller based on $\ell = 8$, the values of the 2 to 8-step-ahead predictions given information up to $t = 100$ (i.e., $y_{102|100}, y_{103|100}, \ldots, y_{108|100}$) are shown in Figure 7.11. Also shown in Figure 7.11 are the likelihood intervals based on τ_{pred} and intervals based on the "prediction error", where these limits are:

$$\hat{y}_{t+k|t} \pm t(n-p;\alpha/2)\sigma_p^2 \tag{7.85}$$

The likelihood intervals based on the standard expression for τ (i.e. the likelihood intervals for the mean values of the k-step-ahead predictions) are not shown on this figure. A comparison of the likelihood intervals based on τ and τ_{pred} is given in the discussion relating to Figures 7.13 and 7.14, the profile t plots for the mean value and a new observation of the 2-step-ahead prediction for the case where $\ell = 4$.

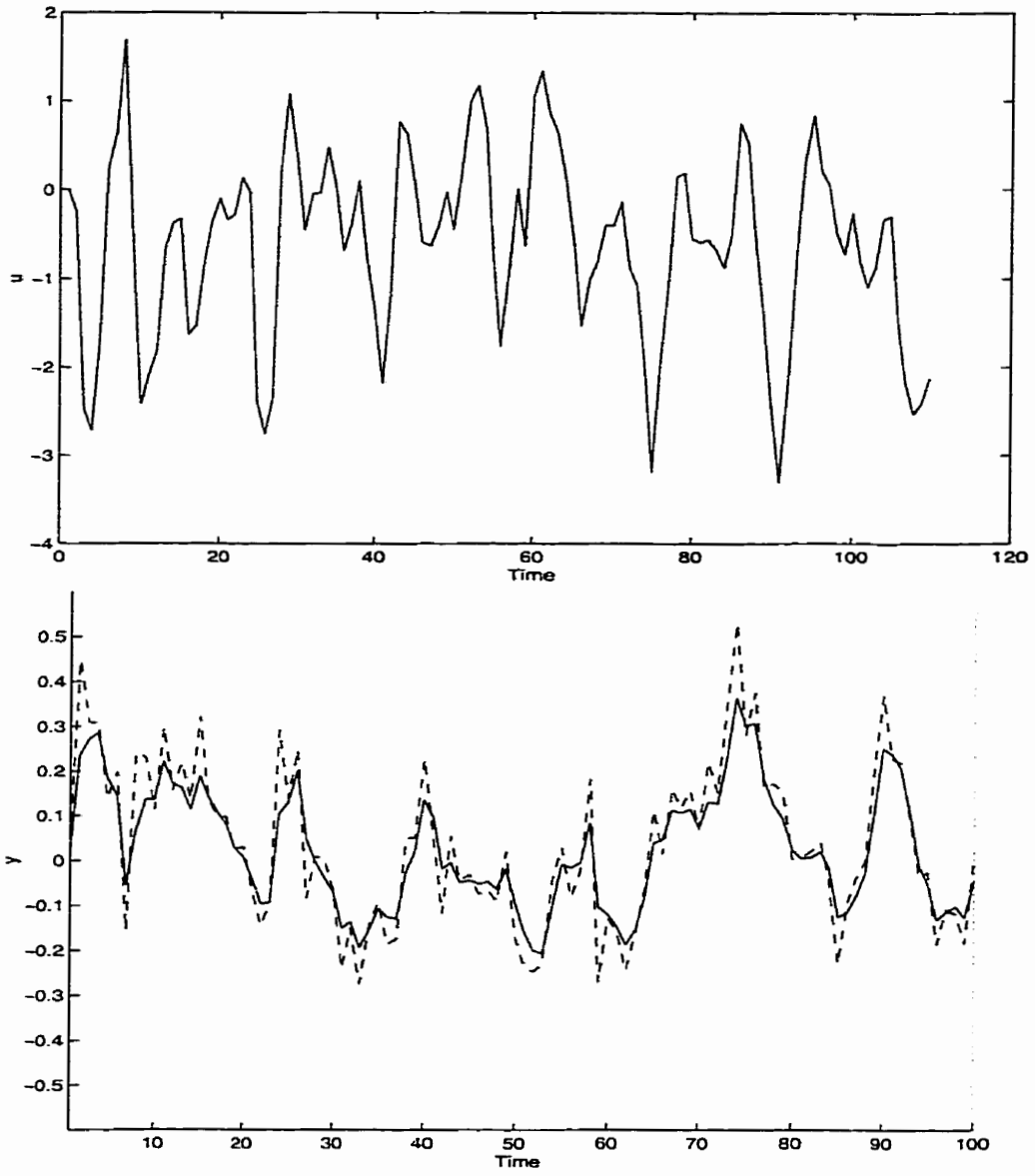

Figure 7.9: Closed-loop simulated data for the case where $\ell = 4$. Key: – the true system; - - the system simulated based on the MLEs of the parameters.

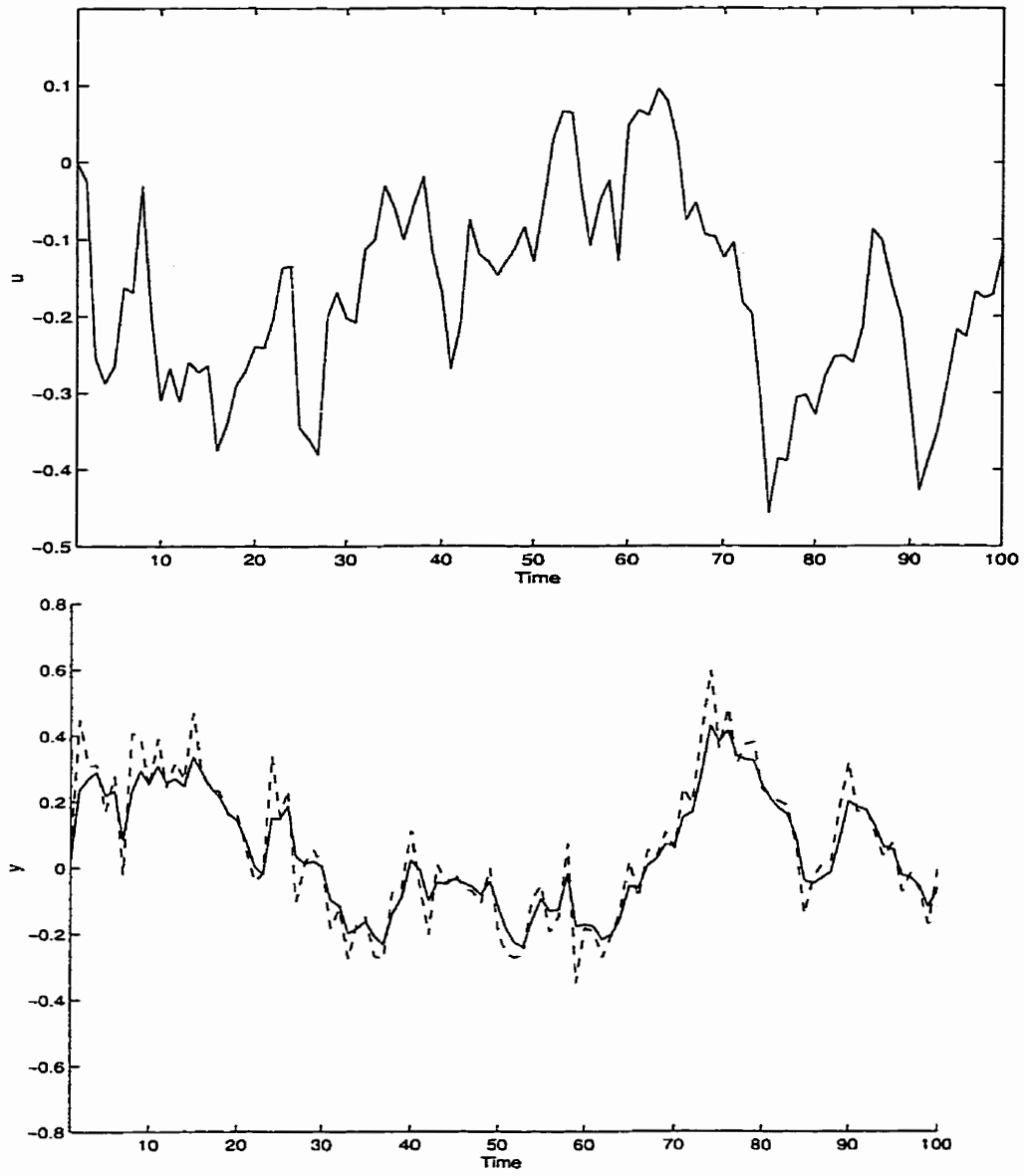

Figure 7.10: Closed-loop simulated data for the case where $\ell = 8$. Key: – the true system; - - the system simulated based on the MLEs of the parameters.

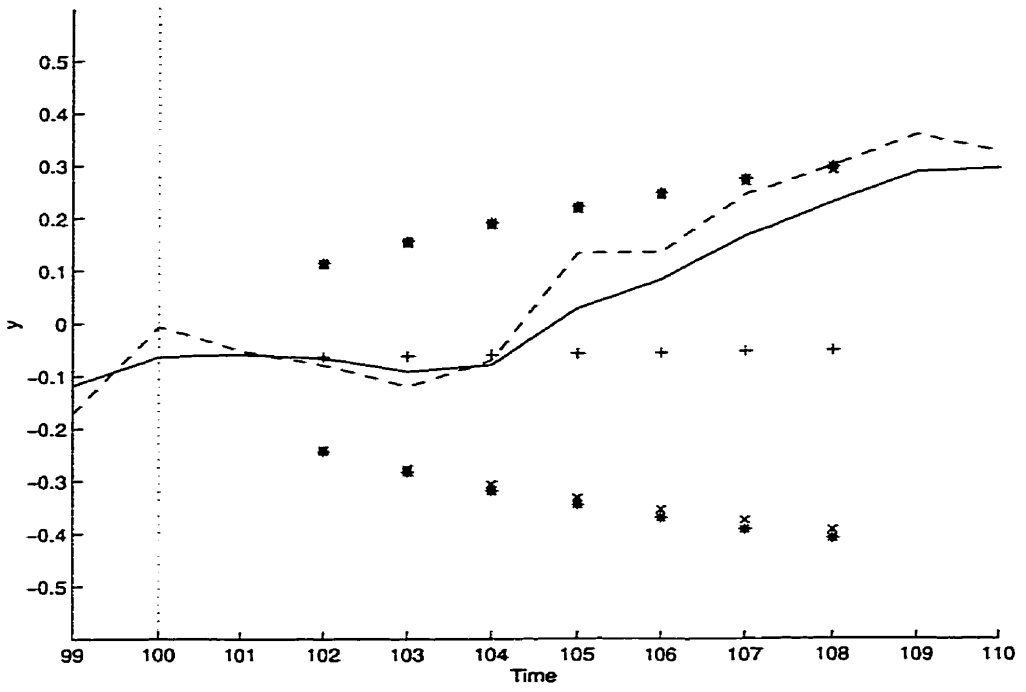

Figure 7.11: Simulated system based $\ell = 8$. Key: – the true system; - - the system simulated based on the MLEs of the parameters; + multi-step-ahead predictions from $t = 100$; × limits of the "prediction error" intervals; $*$ limits of the likelihood intervals for a new prediction.

Although we computed control moves using a random walk model, we based the predictions on the disturbance model $\frac{1}{1-0.99q^{-1}}a_t$; therefore, a slight bias was introduced in the predictions. This is why the 8-step-ahead prediction for the case where $\ell = 8$ and the 4-step-ahead prediction for the case where $\ell = 4$ are not equal to zero. For most of the multi-step-ahead predictions, the "prediction error" bounds are almost coincident with the likelihood intervals, indicating that the uncertainty due to unknown future error dominates the total uncertainty. Typically, the closer a closed-loop system is to being unstable, the larger the discrepancy between the limits of the likelihood intervals and the limits of the "prediction error" limits. The measure of nonlinearity ζ_{min} can be used to judge the relative distance of a system to a stability invertibility boundary. Figure 7.12 shows the results of a simulation of the same system, but with $\ell = 4$. In the two cases considered here both the prediction error limits and the likelihood limits (those that account for uncertainty due to

future random errors) enclose the values of y_t. For cases in which the variance of the disturbance relative to the variance of the input signal is larger than in the current example, the contribution of the parameter uncertainty to the likelihood intervals will be more significant.

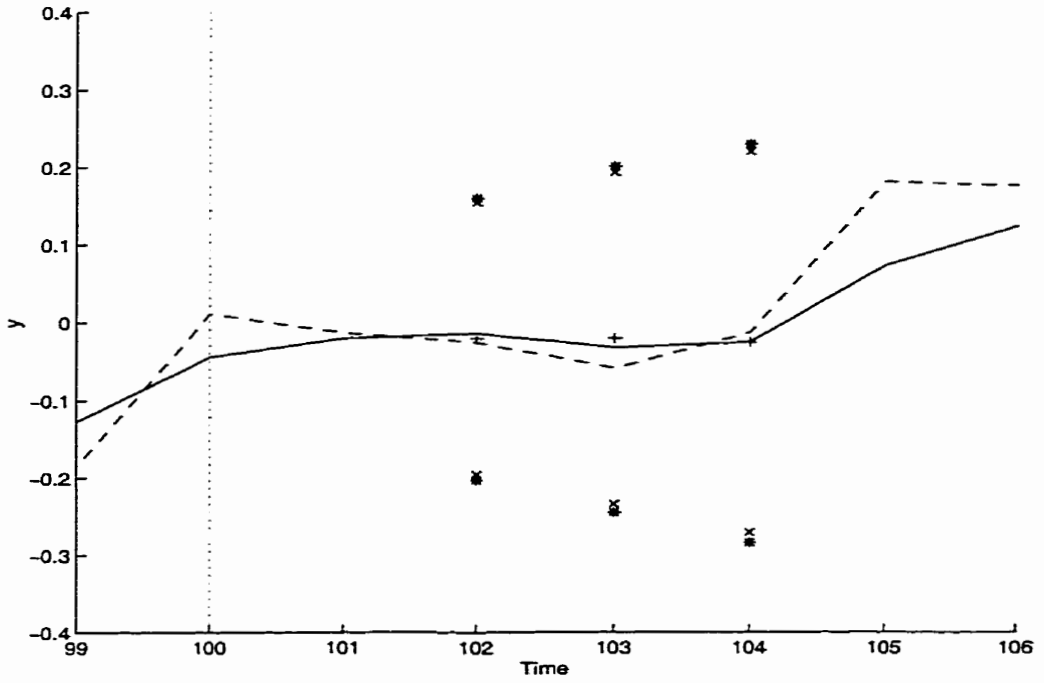

Figure 7.12: Simulated system based $\ell = 4$. Key: – the true system; - - the system simulated based on the MLEs of the parameters; + multi-step-ahead predictions from $t = 100$; × limits of the "prediction error" intervals; * limits of the likelihood intervals for a new prediction.

The profile t plot for the mean value (i.e., the profile t plot based on (7.23)) of the 2-step-ahead prediction $y_{102|100}$ for the case where $\ell = 4$ is shown in Figure 7.13. This prediction shows significant nonlinearity. The profile t plot for a future value of $y_{102|100}$ is shown in Figure 7.14. The likelihood limits at $t = 102$ on Figure 7.12 were obtained from Figure 7.14. Note that the likelihood interval for a future value is significantly wider than the likelihood interval for the mean value and shows almost no nonlinearity. The significance of this is that the error due to unknown future random error is much larger than the error due to parameter uncertainty in this case. Reimer (1995) found that for time series models it is important to consider parameter

uncertainty when the amount of data is very small ($n \leq 50$). We have found that for transfer function models, the nonlinearity of the prediction is strongly affected not only by the amount of data but also by the signal-to-noise ratio, the form of the model and the data themselves. Even when $n = 500$, there are cases for which new multi-step-ahead predictions will show nonlinearity. We advise caution when using

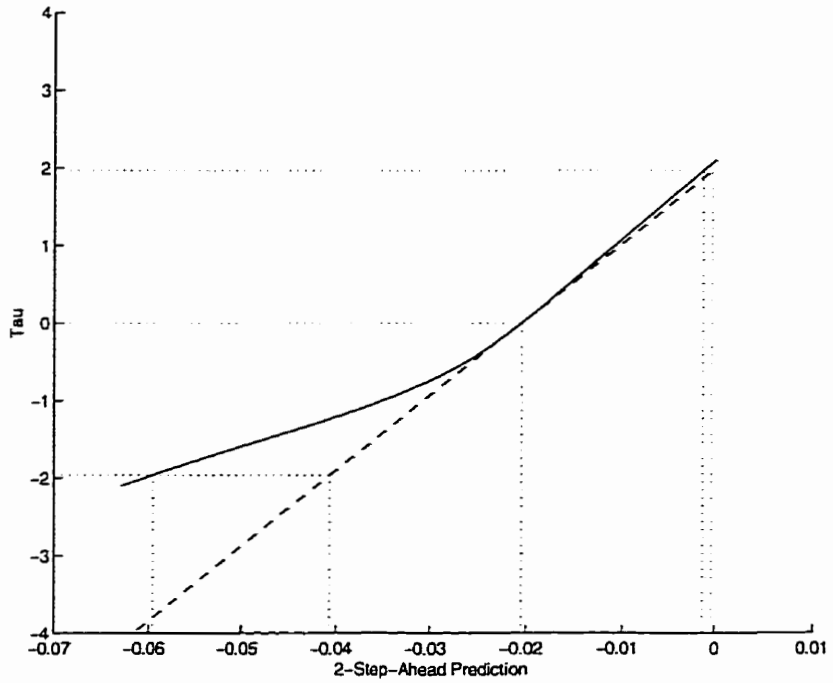

Figure 7.13: Profile t plot for the mean value of the 2-step-ahead prediction given information up to and including $t = 100$ for the case where $\ell = 4$.

linearization confidence intervals even when $g(\boldsymbol{\theta})$ is a prediction. Especially when the mean value of a future prediction is of interest, it is our opinion that the nonlinearity of the inference problem should not be neglected.

7.6.5 Profile Pair Sketching and its Application to Nyquist Plots

A Nyquist plot is a plot of the imaginary component of the frequency response of a dynamic model, $Im[f(e^{i\omega})]$, versus the real component, $Re[f(e^{i\omega})]$. The Nyquist plot

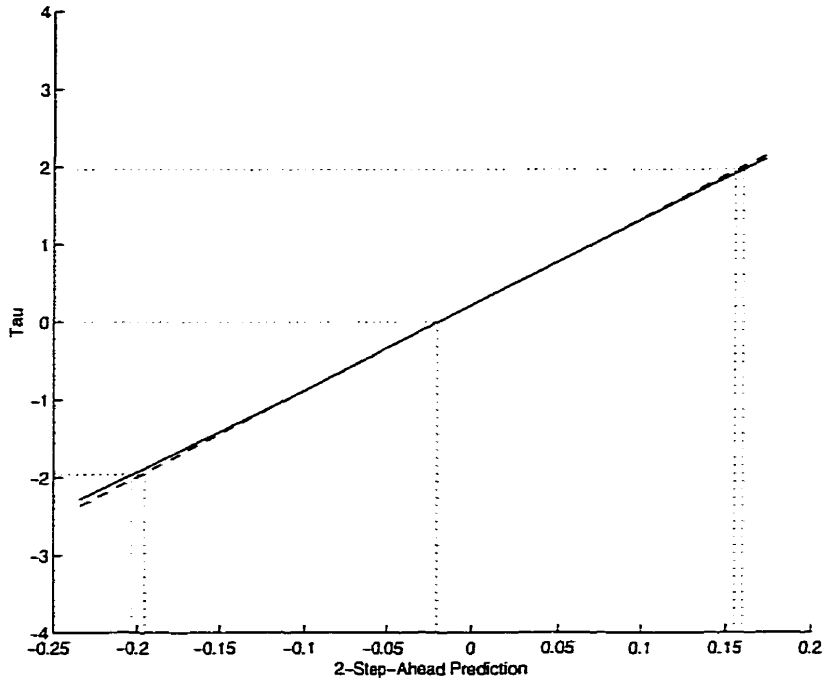

Figure 7.14: Profile t plot for a new observation of the 2-step-ahead prediction given information up to and including $t = 100$ for the case where $\ell = 4$.

is an important tool for examining the behavior and stability of processes (Seborg et al., 1989). In practice, Nyquist plots are based on estimated process models and, therefore, are themselves uncertain. In this section we propose a method for sketching likelihood *regions* about points on an estimated Nyquist curve.

Up to this point in the paper we have considered only likelihood *intervals* for single functions of parameters. In the case of a Nyquist plot, we consider joint likelihood regions for two functions of parameters, $g_1(\boldsymbol{\theta})$ and $g_2(\boldsymbol{\theta})$, where $g_1(\boldsymbol{\theta}) = Re[f(e^{i\omega})]$ and $g_2(\boldsymbol{\theta}) = Im[f(e^{i\omega})]$.

Bates and Watts (1988) proposed a means of using the information gathered throughout the process of profiling to sketch joint likelihood regions for pairs of parameters in nonlinear regression models. We have adapted this method in order to sketch pseudo-ellipses around points on Nyquist plots.

When profiling, it is important to store not only the values of $g(\boldsymbol{\theta})$ and their associated values of τ, but also the results of the constrained optimization problems

$\tilde{\boldsymbol{\theta}}$. A profile trace plot for $g_1(\boldsymbol{\theta})$ and $g_2(\boldsymbol{\theta})$ is a plot of $\tilde{g}_2$ versus $g_1(\boldsymbol{\theta})$, and $g_2(\boldsymbol{\theta})$ versus $\tilde{g}_1$, where $\tilde{g}_1$ is the value of $g_1(\boldsymbol{\theta})$ at the location of the constrained optimum for each iteration of the profiling of $g_2(\boldsymbol{\theta})$, and $\tilde{g}_2$ is the value of $g_2(\boldsymbol{\theta})$ at the constrained optima found over the course of profiling $g_1(\boldsymbol{\theta})$. If r different functions of parameters are being considered, then profile pair sketches may be generated as $g_j(\boldsymbol{\theta})$ versus $g_k(\boldsymbol{\theta})$, for $j, k = 1 \ldots r$, $j \neq k$ (Sulieman, 1998). A profile trace plot provides information about the nonlinearity of $g_j(\boldsymbol{\theta})$ with respect to $g_k(\boldsymbol{\theta})$. For a linear model and linear functions of parameters, the profile traces are straight lines, and the angle of intersection represents the correlation between $g_j(\boldsymbol{\theta})$ and $g_k(\boldsymbol{\theta})$. Profile traces are perpendicular to each other when the parameters or functions of parameters are uncorrelated. For nonlinear models, the profile traces are not necessarily straight lines, and the degree of curvature indicates the degree of nonlinearity of the solution surface in the directions defined by changes in $g(\boldsymbol{\theta})$.

Profile traces are the basis for producing profile pair sketches. Profile pair sketches are a computationally economical way of representing pairwise joint likelihood regions using the information obtained through profiling (Bates and Watts, 1988). It is straightforward to identify the points on the traces for which $\tau = \pm\sqrt{rF(r, n-p, \alpha/2)}$, where r is the number of functions of parameters of interest (Donaldson and Schnabel, 1987). There are four such points (two on each profile trace). It is known that a tangent line to the joint confidence region must be vertical at the points located on the trace of $\tilde{g}_2$ versus g_1, and horizontal at the critical points located on the trace of g_2 versus $\tilde{g}_1$ (Murdoch, 1995). Based on this information alone, and using the parametric description of an ellipse, it is possible to sketch joint likelihood regions. See Figure 7.15 for an explicit algorithm.

The algorithm is based on the fact that after transforming the four points from the parameter space to the τ space, a plot of $\cos^{-1}(\tau_1)$ versus $\cos^{-1}(\tau_2)$ is approximately a straight line, where τ_1 represents the first coordinate of the four tangent points, and

τ_2 represents the second coordinate. To obtain asymmetrical regions which capture the nonlinearity of the inference problem, a cubic spline is fitted through the pairs of points and used to generate a whole series of points which, when transformed back to the parameter space, will outline a pseudo-ellipse.

Figure 7.16 shows the Nyquist plot for model (7.52) based on both the true values of the parameters and the estimated values of the parameters. The regions sketched in Figure 7.16 are easier to generate, in terms of computational effort, than those based on resampling methods and those based on grid-wise contouring methods (Shirt et al., 1994). Despite being based on only four points, these sketches provide good qualitative evidence of the degree of joint uncertainty and the degree of nonlinearity of this example. Note that for this example, the essential features of the uncertainty regions around points on the Nyquist plot are captured by the linearization regions.

The important features of the sketching algorithm are reflected in Figure 7.16. The sketched regions are based on the true nonlinear model and the regions remain true to any implicit constraints on the functions of parameters. For example, when $\omega = 0$, $Im[f(e^{i\omega}]$ always equals zero. Therefore, as ω approaches zero, the uncertainty in $g_2(\boldsymbol{\theta})$ should continue to decrease until at $\omega = 0$ the uncertainty region reduces to an interval. This trend can be seen on Figure 7.16. Also, notice that the overall degree of uncertainty in the estimates of $Re[f(e^{i\omega})]$ and $Im[f(e^{i\omega})]$ decreases when ω gets large, as would be expected, since the frequency response becomes almost independent of the parameter values at large ω.

1. Profile $g_1(\boldsymbol{\theta})$.

2. Over the course of profiling $g_1(\boldsymbol{\theta})$, construct the matrix ${}^1\boldsymbol{M}$, where each row of ${}^1\boldsymbol{M}$ contains the results of the constrained optimization problem solved at each iteration of the profiling algorithm. Each column of the matrix contains values of one of the functions of parameters, one of the parameters, or τ, at each iteration. In table form, ${}^1\boldsymbol{M}$ is

Iteration	τ	$g_1(\bar{\boldsymbol{\theta}})$	---	$g_k(\bar{\boldsymbol{\theta}})$	$\bar{\theta}_1$	---	$\bar{\theta}_p$
1	${}^1m_{1,1}$	${}^1m_{1,2}$	---	${}^1m_{1,k+2}$	${}^1m_{1,k+3}$	---	${}^1m_{1,k+p+2}$
⋮							
${}^1h$	${}^1m_{{}^1h,1}$	${}^1m_{{}^1h,2}$	---	${}^1m_{{}^1h,k+2}$	${}^1m_{{}^1h,k+3}$	---	${}^1m_{{}^1h,k+p+2}$

   where $g_k(\boldsymbol{\theta})$ is the k^{th} function of parameters of interest, $\bar{\theta}_i$ is the i^{th} parameter of the vector of parameter values $\bar{\boldsymbol{\theta}}$, ${}^1h$ is the number of iterations needed to profile $g_1(\boldsymbol{\theta})$, and ${}^1m_{i,j}$ is the ij^{th} element of ${}^1\boldsymbol{M}$

3. Profile $g_2(\boldsymbol{\theta})$.

4. Over the course of profiling $g_2(\boldsymbol{\theta})$, construct the matrix ${}^2\boldsymbol{M}$ as for ${}^1\boldsymbol{M}$ in Step 2.

5. Using the data in ${}^1\boldsymbol{M}$, fit a spline curve $g_{\theta\tau,1}$ to τ as a function of $g_1(\boldsymbol{\theta})$. Also fit a spline $g_{\tau\theta,1}$ to $g_1(\boldsymbol{\theta})$ as a function of τ.

6. Using the data in ${}^2\boldsymbol{M}$, fit a spline curve $g_{\theta\tau,2}$ to τ as a function of $g_2(\boldsymbol{\theta})$. Also fit a spline $g_{\tau\theta,2}$ to $g_2(\boldsymbol{\theta})$ as a function of τ.

7. Use $g_{\theta\tau,1}$ to convert the $g_1(\bar{\boldsymbol{\theta}})$ column of ${}^2\boldsymbol{M}$ to a vector of τ values called τ_{12}.

8. Fit a spline $g_{\tau\tau,2}$ to τ_{12} as a function of the τ data from ${}^2\boldsymbol{M}$.

9. Use $g_{\theta\tau,2}$ to convert the $g_2(\bar{\boldsymbol{\theta}})$ column of ${}^1\boldsymbol{M}$ to a vector of τ values called τ_{21}.

10. Fit a spline $g_{\tau\tau,1}$ to τ_{21} as a function of the τ data from ${}^1\boldsymbol{M}$.

11. Use $g_{\tau\tau,1}$ to compute $q1$, the value of the spline at $\sqrt{rF(r,p-r;\alpha)}$, where r is the number of functions of parameters being considered jointly.

12. Use $g_{\tau\tau,1}$ to compute $q2$, the value of the spline at $-\sqrt{rF(r,p-r;\alpha)}$.

13. Use $g_{\tau\tau,2}$ to compute $p1$, the value of the spline at $\sqrt{rF(r,p-r;\alpha)}$.

14. Use $g_{\tau\tau,2}$ to compute $p2$, the value of the spline at $-\sqrt{rF(r,p-r;\alpha)}$. CONTINUED ON NEXT PAGE

Figure 7.15: A step-by-step algorithm for sketching a profile pair plot for $g_1(\boldsymbol{\theta})$ and $g_2(\boldsymbol{\theta})$ (CONTINUED ON NEXT PAGE).

15. Let

$$\boldsymbol{sp} = \begin{bmatrix} 0 \\ \pi \\ acos\left(\frac{p1}{r\sqrt{F(r,p-r;\alpha)}}\right) \\ acos\left(\frac{p2}{r\sqrt{F(r,p-r;\alpha)}}\right) \end{bmatrix}$$

16. Let

$$\boldsymbol{sq} = \begin{bmatrix} acos\left(\frac{q1}{r\sqrt{F(r,p-r;\alpha)}}\right) \\ acos\left(\frac{q2}{r\sqrt{F(r,p-r;\alpha)}}\right) \\ 0 \\ \pi \end{bmatrix}$$

17. Let $\boldsymbol{a} = \frac{\boldsymbol{sp}+\boldsymbol{sq}}{2}$.

18. Let $\boldsymbol{d} = \boldsymbol{sp} + \boldsymbol{sq}$.

19. If any element of $\boldsymbol{d}$ is negative, change the sign of that element and the sign of the corresponding element of $\boldsymbol{a}$.

20. Let $\boldsymbol{a}^T = [\boldsymbol{a}^T - 2\pi, \boldsymbol{a}^T, \boldsymbol{a}^T + 2\pi]$.

21. Let $\boldsymbol{d}^T = [\boldsymbol{d}^T, \boldsymbol{d}^T, \boldsymbol{d}^T]$.

22. Let $\boldsymbol{S1} = \boldsymbol{a} + \boldsymbol{d}/2$.

23. Let $\boldsymbol{S2} = \boldsymbol{a} - \boldsymbol{d}/2$.

24. Fit a spline $g_{S2,S1}$ to $\boldsymbol{S2}$ as a function of $\boldsymbol{S1}$.

25. Choose a series of 100 equally spaced values from 0 to 2π. Let this vector of values be $\boldsymbol{sps}$.

26. Use $g_{S2,S1}$ to compute the vector $\boldsymbol{sqs}$ which corresponds to the values in $\boldsymbol{sps}$.

27. Let $\boldsymbol{\tau}_{\bullet 1} = cos(\boldsymbol{sps})\sqrt{rF(r,p-r,\alpha)}$.

28. Let $\boldsymbol{\tau}_{\bullet 2} = cos(\boldsymbol{sqs})\sqrt{rF(r,p-r,\alpha)}$.

29. Use $g_{\tau\theta,1}$ to convert $\boldsymbol{\tau}_{\bullet,1}$ to the vector $\boldsymbol{\theta}_{\bullet,1}$.

30. Use $g_{\tau\theta,2}$ to convert $\boldsymbol{\tau}_{\bullet,2}$ to the vector $\boldsymbol{\theta}_{\bullet,2}$.

31. Plot $\boldsymbol{\tau}_{\bullet,2}$ versus $\boldsymbol{\tau}_{\bullet,1}$.

Figure 7.15: A step-by-step algorithm for sketching a profile pair plot for $g_1(\boldsymbol{\theta})$ and $g_2(\boldsymbol{\theta})$ (adapted from Bates and Watts (1988)).

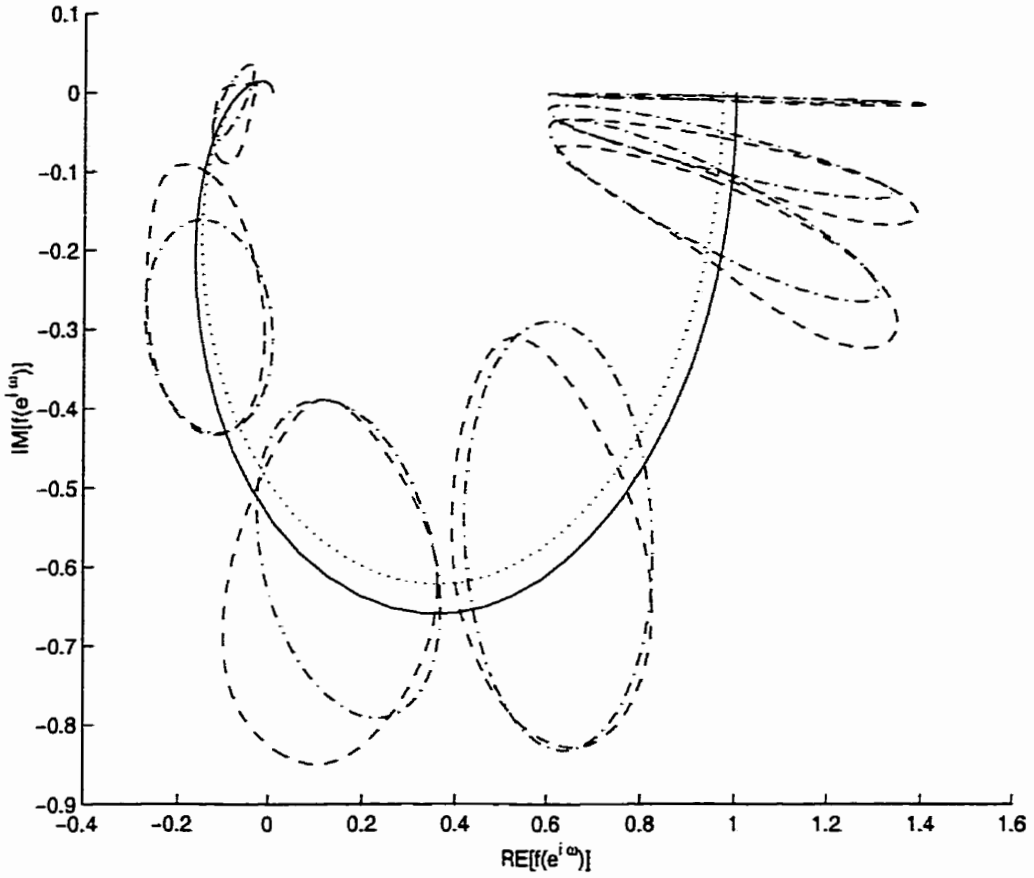

Figure 7.16: Nyquist plot for model (7.52). The pseudo-ellipses were sketched on the basis of profiling data. Key: – true Nyquist curve; ··· Nyquist curve based on the estimated process model; - - joint likelihood regions; .- linearization confidence regions.

7.7 Exact versus Approximate Likelihood Estimation

All of the results reported so far were computed using the algorithm for exact likelihood estimation given in Figure 7.2. Earlier, we introduced algorithms for computing approximate likelihood values. The estimates of the parameters of model (7.52) using the approximate likelihood algorithm are shown in Table 7.2. The profile t plot for f_1 based on these estimation results is shown in Figure 7.17.

Table 7.2: Table of Estimation Results Based on Three Different Estimation Algorithms.

Parameter	True Value	MLE	Approx. MLE	Conditional MLE
b_1	0.0172	0.0199	0.0177	0.0192
d_1	-0.9	-0.8878	-0.8810	-0.8850
f_1	-1.7236	-1.6783	-1.7044	-1.6905
f_2	0.7408	0.6996	0.7227	0.7104

Comparing these results to those presented in Table 7.1 and Figure 7.3, it can be seen that results based on the approximate likelihood algorithm agree quite well with those based on the exact likelihood. The savings, in terms of computation time, using the approximate likelihood algorithm are considerable. Using a Sun Sparc Ultra 1 workstation and MATLAB^{TM} v. 5.1 (Mathworks, 1996), the time required to obtain converged estimates of the parameters of the model and to construct profile t plots for all four parameters using the exact likelihood algorithm was 2.2 hours. To perform the same task using the approximate likelihood algorithm required only 6.4 minutes, and using the the conditional algorithms the results were obtained in 17.2 minutes. However, when a system is expected to be significantly nonlinear, we also expect the differences among the inference results generated by the three algorithms to be more significant. The nonlinearity of a parameter or function of parameters is a complicated function of the model, its parameterization, the amount of data and the data themselves. We recommend computing the measure of nonlinearity ζ_{min} (see Chapter 6) to predict the nonlinearity of $g(\boldsymbol{\theta})$. Only when the nonlinearity is expected to be high is it necessary to use the computationally expensive full maximum likelihood algorithm.

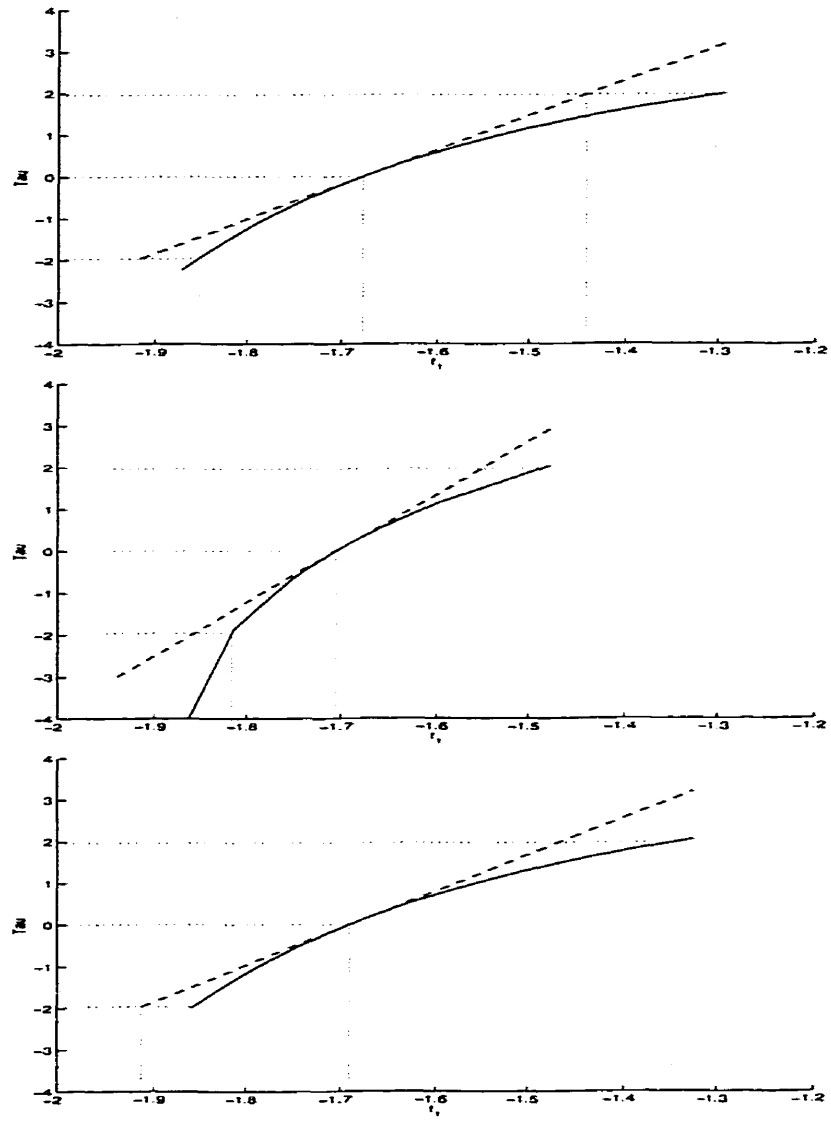

Figure 7.17: Profile t plots for parameter f_1 based on the exact, approximate and conditional maximum likelihood algorithms, respectively.

7.8 Conclusions

Profiling has been employed to estimate likelihood intervals for functions of parameters commonly used in control. Although some functions of parameters behaved quite linearly, and some linearization confidence intervals were good approximations of the corresponding likelihood intervals, others were not. The purpose of this work was to demonstrate the methodology and assess its merits and limitations. The cases considered here suggest that care should be taken when employing linearization inference

results, and that there is a place for more computationally intensive but more reliable approaches to inference. A broad simulation study would be worthwhile to confirm this conjecture.

It is important to keep in mind that the nonlinearity of a function of parameters is influenced by a range of factors, including the form of the model, its parameterization, the proximity of the parameter vector to a stability/invertibility boundary, the amount of data, and the data themselves. In many control applications, data sets of more that 500 observations are common, and the sheer quantity of data has been used as reason enough to assume linearity since the linearization inference results are asymptotically exact as n approaches infinity (Ljung, 1987). However, the amount of data required for the asymptotic results to be reliable is poorly known. Also, the amount of data needed depends on the model, its parameterization, the values of the parameters and the signal-to-noise ratio. In the face of all of these factors, we recommend generalized profiling, together with the measure of nonlinearity introduced in Chapter 6, as means by which to ensure reliable inference results. We see that generalized profiling has application in design of experiments where it could be used in determining the length of an experiment, or in comparative evaluations of competing experimental designs (see Chapter 5 and Shirt, 1997). There are also important issues related to the use of profiling in adaptive control applications which have not yet been addressed. It may be interesting to track how error propagates differently depending on whether direct or indirect synthesis of the controller is used.

The theory and methods presented here have been illustrated for cases for which there was no model uncertainty; however, in its most general form, profiling is a likelihood ratio method and there is no inherent restriction that the form of the model defining the likelihood for the numerator of the ratio be the same as that for the denominator. This implies that the extension of the method to cases of undermodeling or overmodeling would be straightforward.

7.9 Acknowledgments

The authors gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada and the School of Graduate Studies of Queen's University.

7.10 Nomenclature

a_t	= white noise sequence
$a_1, a_2, \ldots, a_{na}$	= coefficients of the $A(q^{-1})$ polynomial
$A(q^{-1})$	= polynomial in the backshift operator q^{-1}
$b_1, b_2, \ldots, b_{nab}$	= coefficients of the $B(q^{-1})$ polynomial
$B(q^{-1})$	= polynomial in the backshift operator q^{-1}
c	= a constant
$c_1, c_2, \ldots, c_{nc}$	= coefficients of the $C(q^{-1})$ polynomial
$cov(\hat{\boldsymbol{\theta}})$	= variance covariance matrix of $\hat{\boldsymbol{\theta}}$
$C(q^{-1})$	= polynomial in the backshift operator q^{-1}
d	= delay between a change in u_t and its effect on y_t
$d_1, d_2, \ldots, d_{nd}$	= coefficients of the $D(q^{-1})$ polynomial
$D(q^{-1})$	= polynomial in the backshift operator q^{-1}
$e_{t+k\|t}$	= prediction error due to unknown future random errors
$f_1, f_2, \ldots, f_{nf}$	= coefficients of the $F(q^{-1})$ polynomial
$F(p, n-p; \alpha)$	= upper α quantile for the F distribution with p and $n-p$ degrees of freedom
$F(q^{-1})$	= polynomial in the backshift operator q^{-1}

$g(\boldsymbol{\theta})$	= a function of parameters
$gain$	= steady state gain of the process
$G(q^{-1})$	= process transfer function
$G_c(q^{-1})$	= controller transfer function
$G_{OL}(q^{-1})$	= open loop transfer function
$G_p(q^{-1})$	= process transfer function
$\boldsymbol{h}$	= $p \times 1$ vector of constants
$H(q^{-1})$	= disturbance transfer function
$IM[f(e^{i\omega})]$	= imaginary component of the frequency response of the model
k	= the number of sampling intervals into the future for which a prediction of y_t is to be made
ℓ	= prediction horizon
$L(q^{-1})$	= polynomial in the backshift operator q^{-1}
$L(\boldsymbol{\theta})$	= likelihood function evaluated at $\boldsymbol{\theta}$
$\mathcal{L}(\boldsymbol{\theta})$	= natural logarithm of the likelihood function of $\boldsymbol{\theta}$
$LI(g(\boldsymbol{\theta}))$	= likelihood interval for $g(\boldsymbol{\theta})$
LR	= likelihood ratio
$\mathcal{LR}$	= natural logarithm of the likelihood ratio
$M(q^{-1})$	= polynomial in the backshift operator q^{-1}
n	= number of observations
na, nb, nc, nd, nf	= orders of the polynomials $A(q^{-1})$, $B(q^{-1})$, $C(q^{-1})$, $D(q^{-1})$, and $F(q^{-1})$, respectively
$N(q^{-1})$	= polynomial in the backshift operator q^{-1}
p	= number of estimated parameters
$P(q^{-1})$	= polynomial of degree $k-1$ resulting from the Diophantine decomposition

q^{-1} = backshift operator

$Q(q^{-1})$ = polynomial resulting from the Diophantine decomposition

$RE[f(e^{i\omega})]$ = real component of the frequency response of the model

s = estimated standard deviation of the random errors

$S(\boldsymbol{\theta})$ = $\boldsymbol{y_n}'\boldsymbol{\Omega}_n^{-1}\boldsymbol{y_n}$

se = standard error

$t(n-p;\alpha/2)$ = upper $\alpha/2$ quantile of the t distribution with $n-p$ degrees of freedom

$\boldsymbol{u}$ = vector of inputs

u_f = filtered inputs

u_t = input to the process at time t

$\boldsymbol{V}$ = $n \times p$ matrix of elements v_{ij} representing the first derivative of $f(\boldsymbol{x}_i, \boldsymbol{\theta})$ with respect to the j^{th} parameter

w_t = a time series

$\boldsymbol{x}$ = $1 \times m$ row vector of m independent variables

$\boldsymbol{X}$ = $n \times p$ matrix of elements x_{ij} representing the level of the j^{th} independent variable for observation i

y = response variable

$\hat{y}_{t+k|t}$ = predicted value of $\hat{y}_{t+k}$ given information up to time t

$\boldsymbol{y}$ = $n \times 1$ column vector of values of the response variable

z_t = a time series transformed using Ansley's transformation

Greek letters

α = significance level

γ = constant chosen based on the desired level of confidence and the variance of the additive white noise

$\delta(g(\boldsymbol{\theta}))$	= studentized value of $g(\boldsymbol{\theta})$
Δ	= backward difference operator
$\boldsymbol{\epsilon}$	= $n \times 1$ column vector of random errors
θ_i	= i^{th} parameter of a model
$\boldsymbol{\theta}$	= $p \times 1$ vector of parameters
$\hat{\boldsymbol{\theta}}$	= $p \times 1$ vector of maximum likelihood estimates of the parameters
$\tilde{\boldsymbol{\theta}}$	= location of a constrained maximum of $L(\boldsymbol{\theta})$
σ_a^2	= variance of the white noise sequence a_t
σ_p^2	= variance of the prediction error $e_{t+k\|t}$
$\tau(g(\boldsymbol{\theta}))$	= profile t statistic for $g(\boldsymbol{\theta})$
τ_{pred}	= profile t statistic for new predictions
$\chi^2(1)$	= the chi-squared distribution with 1 degree of freedom
$\sigma_a^2\boldsymbol{\Omega}$	= $n \times n$ covariance matrix for $\boldsymbol{y}$

Superscripts

$*$	= a true value
$\hat{}$	= a maximum likelihood estimate
$\tilde{}$	= a constrained estimate

Abbreviations

AR	autoregressive
ARMA	autoregressive moving average
ARMAX	autoregressive moving average with exogenous inputs
ARX	autoregressive with exogenous inputs
FIR	finite impulse response

iid	independently and identically distributed
MLE	maximum likelihood estimate
SISO	single input single output

Chapter 8

Conclusion

In this chapter, a summary is made of the conclusions reached in each of the five manuscripts. The focus is on clearly identifying the contributions made to chemical engineering and statistical theory and practice. In general, we made theoretical contributions by:

1. developing a geometrical treatment of inference about a function of parameters $g(\boldsymbol{\theta})$, and identifying cases for which profiling fails (Chapters 3 and 4);

2. showing the equivalence of two approaches to generalized profiling (Chapters 3 and 4);

3. developing an expression for the expected value of the likelihood ratio function for time series models (Chapter 5);

4. extending the generalized algorithm to discrete transfer functions - a class of models often used in control applications (Chapter 7), and

5. deriving an expression for the τ profiling statistic which takes into account both parameter uncertainty and uncertainty due to unknown future errors (Chapter 7).

In general, we made practical contributions by:

1. identifying cases in control for which profiling is not appropriate, and proposing an alternate method for one such case(Chapters 3 and 4);

2. showing that functions of parameters of interest in control may sometimes behave nonlinearly, and that generalized profiling is a valuable tool by which to obtain likelihood intervals in these cases (Chapter 7);

3. developing a technique for judging *a priori* the expected nonlinearity of the parameters and functions of parameters of ARMA models, and demonstrating the usefulness of this tool in designing the length of an experiment (Chapter 5);

4. developing a new empirical measure of nonlinearity for ARMA models (Chapter 6);

5. developing a pseudo-profiling technique for quickly judging, in a qualitative way, the nonlinearity of a function of parameters (Chapter 6), and

6. adapting a computationally efficient methodology for sketching likelihood regions for use in sketching uncertainty regions around points on Nyquist plots (Chapter 7).

8.1 Contributions to Theory

Explicit development of the generalized profiling algorithm from both reparameterization and optimization perspectives, and a proof of their equivalence, has helped to elucidate the merits and limitations of this method. This enhanced understanding of generalized profiling has helped to identify the scope of its applicability and usefulness. One important finding was that if $g(\boldsymbol{\theta})$ reaches an unconstrained optimum at $\hat{\boldsymbol{\theta}}$, the generalized profiling algorithm fails. An expression for the expected value of the

square of the profiling statistic τ was developed for the case of an ARMA model. This is an important contribution because it can be used to gage the expected nonlinearity of a function of parameters as a function of n. Furthermore, the limits of the expected likelihood intervals can be compared to the linearization results to draw conclusions about how much data is needed for asymptotic inference results to apply. Another important consequence of this development is that it may provide a basis upon which to judge the contribution of anomalies in a data set to the overall nonlinearity of a function of parameters.

Extension of the work of Bates and Watts (1988), Lam and Watts (1991), Chen (1991), and Chen and Jennrich (1996) to utilize the profiling algorithm for functions of parameters of transfer function models was straightforward. From a theoretical perspective, the interesting aspects of using profiling in the context of transfer function models are related to special issues involving some of the functions of parameters we have considered, namely multi-step-ahead predictions and gain margins. For the case of predictions, a modified expression for the τ statistic was developed so that profiling could be used to construct likelihood intervals that would account for both parametric uncertainty and uncertainty due to unknown future random errors. For profiling the gain margin, it was found to be important to define the problem and identify what information is desired. Associated issues were discussed, and three cases were examined. The case for which the values of the parameters in the controller are fixed was discussed in detail. Consideration of these functions of parameters motivated new extensions of the profiling algorithm. The issues involved in profiling multi-step-ahead predictions and gain margins are common to many other functions of parameters. From our look at two specific examples comes a methodology for solving similar problems.

8.2 Contributions to Practice

Much of the value in the development of both the theory and application of the generalized profiling algorithm is in making this statistical method understandable and available to engineers. To some extent, the statistical literature is impenetrable to engineers, thereby presenting a barrier to the use of new statistical methods, and creating a gap between theory and practice. Hopefully, this thesis has helped to bridge the gap between theory and application of likelihood intervals.

The analysis of the reparameterization and optimization approaches to generalized profiling allowed us to identify the limitations of profiling. From a practical point of view, this is important since some functions of interest, such as the measure of controller performance Δ_{perf}, may reach an optimum at $\hat{\boldsymbol{\theta}}$. In these cases profiling fails, but the minimization/maximization approach to finding likelihood intervals can be used.

There is a growing body of literature on measuring uncertainty in transfer function models which has been motivated by the needs and demands of robust control theory. The focus has been on developing hard error bounds, although there have been several papers advocating a shift in paradigm towards soft (probabilistic) error bounds (Goodwin et al., 1992; DeVries and van den Hof, 1994; Pintelon, 1994). We have shown that functions of parameters used by control engineers may behave nonlinearly. In light of this, generalized profiling is a more reliable and appropriate method for making inferences about functions of parameters than the linearization method. For cases involving two functions of parameters considered jointly, it is appropriate to construct likelihood regions. The algorithm for sketching likelihood regions proposed by Bates and Watts (1988) has been adapted for use with functions of parameters of transfer function models. The algorithm was then used to construct likelihood regions around points on Nyquist plots in a way that is more computationally efficient than other methods proposed in the literature.

Expected profiling has been developed for use in designing the length of an experiment when the purpose of that experiment is to estimate a time series model. This fills a gap in the practice of designing dynamic experiments. Also, for cases in which the data have already been collected, expected profiling is an important means of evaluating whether commonly used asymptotic results are appropriate for the given model and information content of the available data.

The new measure of nonlinearity proposed for ARMA models is a "quick and easy" measure intended to be easy to implement and easy to interpret. The measure developed by Ravishanker (1994), based on the work of Bates and Watts (1980), requires expressions for the first and second derivatives of the model with respect to the parameters, and involves performing operations on three-dimensional arrays. The effort involved in computing this measure of curvature is a barrier to its use. The measure developed in this thesis is intended to be computationally accessible to chemical engineers so that it may be more attractive for use in practice. Similarly, a pseudo-profiling algorithm has been proposed for use with a broad class of models as a means to quickly and easily judge, in a qualitative way, the nonlinearity of a function of parameters.

It may be said that any model, and any statistic derived therefrom, is virtually useless without an associated measure of uncertainty (adapted from Wahlberg and Ljung, 1992). In this thesis we provide reliable means for estimating and predicting uncertainty in control-relevant statistics.

Chapter 9

Recommendations

In all of the work reported here, it has been assumed that the proposed model is sufficiently complex to capture the true behavior of the system. There is important work left to be done to incorporate model mismatch into the uncertainty methodologies developed in this thesis. As discussed in the literature review, a significant amount of work related to model mismatch has been reported in the literature on robust control. However, most of this work takes a "hard bound" approach to uncertainty. There are important contributions yet to be made in the area of developing statistical uncertainty bounds which are reliable and account for both parameter and model uncertainty.

A superficial review of the methodologies discussed in this thesis suggests that including model uncertainty in generalized profiling and expected profiling would be straightforward. Incorporating model mismatch would amount to solving the optimization problems over a set of possible model forms rather than searching for an optimum based on one "true" form of the model.

The relationship between the curvature of $g(\boldsymbol{\theta})$ and the curvature of $f(\boldsymbol{x}, \boldsymbol{\theta})$ should be investigated further, with a view towards developing a methodology for predicting the nonlinearity of $g(\boldsymbol{\theta})$ given $f(\boldsymbol{x}, \boldsymbol{\theta})$. This idea was inspired by the fact that $g(\boldsymbol{\theta})$

may be linear even when one or more of the parameters is nonlinear, or $g(\boldsymbol{\theta})$ may be nonlinear when all of the parameters are linear. It should be possible, by studying the geometry of $f(\boldsymbol{x}, \boldsymbol{\theta})$ and $g(\boldsymbol{\theta})$, to predict the nonlinearity of $g(\boldsymbol{\theta})$. The recent work of Kang and Rawlings (1998) provides a methodology by which to compute a measure of nonlinearity for a function of parameters. This marginal measure of nonlinearity is based on the Bates and Watts measures and is derived from the perspective of a reparameterization of the model such that $g(\boldsymbol{\theta})$ becomes a parameter of the model. This approach could serve as a starting point from which to interpret the geometry.

We now propose an hypothesis for why a function of parameters (e.g. a model prediction) is often found to behave linearly even when the joint confidence region for $\boldsymbol{\theta}$ is significantly nonlinear. We build on the linearization results presented in Chapter 6.

Recall that for a linear model and a linear function of the parameters $g(\boldsymbol{\theta}) = \boldsymbol{a}^T\boldsymbol{\theta}$, a $(1-\alpha_1)100$ % confidence *interval* for $g(\boldsymbol{\theta})$ includes all values of $g(\boldsymbol{\theta})$ defined over the values of $\boldsymbol{\theta}$ satisfying

$$\frac{S(\boldsymbol{\theta}) - S(\boldsymbol{\theta}_{LS})}{S(\boldsymbol{\theta}_{LS})} \leq \frac{F(1, n-p; \alpha_1)}{n-p} \tag{9.1}$$

where $\boldsymbol{\theta}_{LS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. However, a $(1-\alpha_2)100$ % confidence *region* for $\boldsymbol{\theta}$ is defined based on

$$\frac{S(\boldsymbol{\theta}) - S(\boldsymbol{\theta}_{LS})}{S(\boldsymbol{\theta}_{LS})} \leq \frac{p\, F(p, n-p; \alpha_2)}{n-p} \tag{9.2}$$

The limits of a $(1-\alpha_1)100$ % confidence interval for $g(\boldsymbol{\theta})$ occur at the point where a contour of $g(\boldsymbol{\theta})$ is tangent to the contour defined by (9.1). While this contour is used to define a $(1-\alpha_1)100$ % confidence interval for $g(\boldsymbol{\theta})$, it also defines a confidence region for $\boldsymbol{\theta}$ with a different level of confidence. For example, if $\alpha_1 = 0.01$, $n = 500$ and $p = 2$, a 99 % confidence interval for $g(\boldsymbol{\theta})$ will be tangent to the 95 % confidence

region for $\boldsymbol{\theta}$ since $F(1, 48; 0.01) \approx 2\ F(2, 48; 0.05)$. Since contours of $S(\boldsymbol{\theta})$ behave more linearly as the value of α increases, even though a $(1-\alpha)100$ % confidence region for the parameters behaves nonlinearly, a $(1-\alpha)100$ % interval for $g(\boldsymbol{\theta})$ may behave linearly because it is based on a confidence region for $\boldsymbol{\theta}$ which has a larger value of α.

We have not pursued these ideas further except to note that they do not provide the whole answer. For example, it is common to observe profile t plots for *individual* parameters which are highly nonlinear and profile t plots for model predictions based on those parameters which are linear. In such cases, the effective value of α for the contour of $S(\boldsymbol{\theta})$ which is used to compute inferences for the individual parameters and the predictions is the same. In these cases we suspect that the reasons for the linearity of the predictions involves the orientation of the contours of $g(\boldsymbol{\theta})$ with respect to the contours of $S(\boldsymbol{\theta})$.

With respect to the new tools developed, namely expected profiling and the measures of nonlinearity, there remains work to be done to see how these methods perform over a large number of examples. Specifically for expected profiling, there is important work to be done to generalize the method so that one may consider the case where $\boldsymbol{\theta}^* \neq \hat{\boldsymbol{\theta}}$ (i.e the vector of estimated parameters is not equal to the true values of the parameters), and the case where there is model mismatch. All of the new methods should be extended to include transfer function models.

The work reported here has demonstrated that generalized profiling is a viable and reliable means by which to make inferences about $g(\boldsymbol{\theta})$. The use of profiling to study the effect of design of experiment on the ultimate uncertainty of various control-relevant statistics could constitute an important contribution. This work was begun by Shirt (1997).

When n is large and an exact likelihood algorithm is used to profile functions of parameters of time series or transfer function models, the computation time can be

prohibitively large. There is work to be done to improve the computational efficiency of the algorithm. For the special case of a linear model and a nonlinear function of parameters, we propose the following algorithm for constrained optimization as one way to improve the computational efficiency.

For this special case, the optimization problem to be solved at each iteration of the profiling algorithm is

$$\begin{aligned} \text{Minimize} \quad & J = (\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\theta})^T(\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\theta}) \\ \text{Subject to} \quad & g(\boldsymbol{\theta}) = c \end{aligned} \tag{9.3}$$

We develop the optimization algorithm by linearizing the function $g(\boldsymbol{\theta})$ around a series of values of $\boldsymbol{\theta}$. The linear approximation to $g(\boldsymbol{\theta})$ is $\bar{g} = g(\bar{\boldsymbol{\theta}}) + \boldsymbol{a}^T\Delta\boldsymbol{\theta}$, where $\boldsymbol{a} = \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}}$ and $\Delta\boldsymbol{\theta} = \boldsymbol{\theta} - \bar{\boldsymbol{\theta}}$. Then the location of the solution to the linear optimization problem is:

$$\boldsymbol{\theta} = \boldsymbol{\theta}_{LS} + \lambda(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{a} \tag{9.4}$$

where

$$\lambda = \frac{\boldsymbol{a}^T\boldsymbol{\theta}_{LS} - c}{\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{a}} \tag{9.5}$$

The idea for the new optimization algorithm is based on iteratively updating the value of λ based on linearizations of $g(\boldsymbol{\theta})$ at a succession of points. Define

$$g^{(i)} = g(\boldsymbol{\theta}_i) \tag{9.6}$$

and

$$\bar{g}^{(i)} = g^{(i)} + \left.\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(i)}} \Delta\boldsymbol{\theta}^{(i)} \tag{9.7}$$

where

$$\Delta\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}_{LS} - \boldsymbol{\theta}^{(i)} \tag{9.8}$$

At each iteration of the algorithm we update $\lambda^{(i)}$ using

$$\lambda^{(i)} = \frac{(\boldsymbol{a}^{(i)})^T\boldsymbol{\theta}_{LS} - c}{(\boldsymbol{a}^{(i)})^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{a}^{(i)}} \tag{9.9}$$

where $\boldsymbol{a}^{(i)} = \left.\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(i)}}$. Then compute a new value of $\boldsymbol{\theta}^{(i)}$ using

$$\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}_{LS} + \lambda^{(i)}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{a}^{(i)} \tag{9.10}$$

Continue to compute values of $\lambda^{(i)}$ and $\boldsymbol{\theta}^{(i)}$ until $g(\boldsymbol{\theta}^{(i)}) = c$. This algorithm reduces the constrained p-dimensional optimization to a one-dimensional search for λ, thereby making the solution procedure very quick and easy. We note that it is possible to construct an example involving a highly nonlinear function $g(\boldsymbol{\theta})$ for which this algorithm will converge to a value of $\boldsymbol{\theta}^{(i)}$ which satisfies the constraint but does not give the lowest possible value of $S(\boldsymbol{\theta})$. However, for most forms of $g(\boldsymbol{\theta})$, this algorithm will give reliable results or will at least provide good starting guesses to feed into a more general constrained optimization package.

There remain many interesting applications of profiling to issues in control which have not been addressed. In Chapter 7 we identified that a modified profiling algorithm would have to be employed to profile $G_{OL}(q^{-1}) = G_p(q^{-1})G_c(q^{-1})$ if uncertainty in both $G_p(q^{-1})$ and $G_c(q^{-1})$ were to be taken into account. A similar modification to

the algorithm would likely be required if an adaptive controller were to be considered. It may also be interesting to study differences in the way error propagates from data to model predictions depending on whether a direct or indirect approach to controller design is used (Goodwin and Sin, 1984).

Sulieman (1998) used profiling as the basis for a new approach to sensitivity analysis. Her ideas, along with those presented here, could be used to develop a methodology for identifying which parameters of a model have the greatest affect on the function of parameters of interest. This could have important application in the area of adaptive control, where decisions are made about which parameters to update on-line.

We anticipate that as model-based control strategies become more and more prevalent in industry, statistical analyses will become an integral part of evaluating the merits and limitations of proposed strategies. Generalized profiling has the potential to play an important role in marrying process monitoring schemes with automatic control schemes to achieve a truly integrated approach to process optimization.

There is much work to be done to refine and generalize the statistical algorithms developed in this thesis, and there is potential for further application of these algorithms in almost every sector of chemical engineering.

Bibliography

Abraham, B. and Ledolter, J. (1983), *Statistical Methods for Forecasting*, John Wiley and Sons, New York, NY.

Agrawal, R. (1993), Nonlinear models for mixture experiments, MSc thesis, Queen's University, Department of Mathematics and Statistics.

Akcay, H. and Ninness, B. (1998), 'Rational basis functions for robust identification from frequency and time-domain measurements', *Automatica* **34**, 1101–1117.

Ali, M. (1977), 'Analysis of autoregressive-moving average models: Estimation and prediction', *Biometrika* **64**, 535–545.

Alpen, J. and Gelb, R. (1990), 'Standard errors and confidence intervals in nonlinear regression: Comparison of monte carlo and parametric statistics', *J. Phys. Chem.* **94**, 4747–4751.

Al'tshuler, S. (1983), 'Parameter estimation method for autoregressive-moving average processes', *Automat. Remote Control* **43**, 979–990.

Ansley, C. (1979), 'An algorithm for the exact likelihood of a mixed autoregressive moving average process', *Biometrika* **66**, 59–65.

Ansley, C. and Newbold, P. (1980), 'Finite sample properties of estimators for autoregressive moving average models', *J. Econometrics* **13**, 159–183.

Åström, K. J. (1970), *Introduction to Stochastic Control Theory*, Academic Press, London.

Åström, K. J. (1980), 'Maximum likelihood and prediction error methods', *Automatica* **16**, 551–574.

Åström, K. J. and Wittenmark, B. (1990), *Computer Controlled Systems: Theory and Design, 2nd Ed.*, Prentice-Hall, New Jersey.

Barndorff-Nielson, O. (1986), 'Inference on full and partial parameters based on the standardized signed log likelihood ratio', *Biometrika* **73**, 307–322.

Bates, D. and Watts, D. (1980), 'Relative curvature measures of nonlinearity', *J. R. Statist. Soc. B* **42**, 1–25.

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

Bates, D. and Watts, D. (1991), 'Model building in chemistry using profile t and trace plots', *Chemo. Int. Lab. Sys.* **10**, 107–116.

Beale, E. (1960), 'Confidence regions in non-linear estimation', *J. R. Statist. Soc. B* **22**, 41–76.

Bolviken, E. and Skovlund, E. (1996), 'Confidence intervals from monte carlo tests', *JASA* **91**, 1071–1078.

Box, G. E. P. and Jenkins, G. M. (1976), *Time Series Analysis: Forecasting and Control*, Holden Day, San Francisco.

Box, G. and Tiao, G. (1973), *Baysian Inference in Statistical Analysis*, Addison-Wesley, Reading, MA.

Box, M. (1971), 'Bias in nonlinear estimation', *J. R. Statist. Soc. B* **32**, 171–201.

Canale, M., Malan, S. and Milanese, M. (1998), 'Model quality evaluation in identification for H_∞ control', *IEEE Trans. Autom. Control* **43**, 125–132.

Carr, N. (1960), 'Kinetics of catalytic isomerization of n-pentane', *Ind. Eng. Chem. Res.* **52**, 391–396.

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

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

Chen, J., Nett, C. and Fan, M. (1995), 'Worst case system identification in h_∞: Validation of a priori information, essentially optimal algorithms, and error bounds', *IEEE Trans. Autom. Control* **40**, 1260–1265.

Chui, C. and Chen, G. (1987), *Kalman Filtering with Real-Time Applications*, Springer-Verlag, Berlin, Germany.

Clarke, G. (1987*a*), 'Approximate confidence limits for a parameter function in nonlinear regression', *JASA* **82**, 221–230.

Clarke, G. (1987*b*), 'Marginal curvatures and their usefulness in the analysis of nonlinear regression models', *JASA* **82**, 844–850.

Cook, R. and Goldberg, M. (1986), 'Curvatures for parameter subsets in nonlinear regression', *The Annals of Statistics* **14**, 1399–1418.

Cook, R. and Weisberg, S. (1990), 'Confidence curves in nonlinear regression', *JASA* **85**, 544–551.

Cook, R. and Witmer, J. (1985), 'A note on parameter-effects curvature', *JASA* **80**, 872–878.

Cordeiro, G. G. P. and Botter, D. (1994), 'Improved likelihood ratio tests for dispersion models', *Int. Statist. Review* **62**, 257–274.

Cox, C. and Ma, G. (1995), 'Asymptotic confidence bands for generalized nonlinear regression models', *Biometrics* **51**, 142–150.

Cox, D. and Hinkley, D. (1974), *Theoretical Statistics*, Chapman and Hall, London, England.

Crowe, C. (1976), Personal correspondence. Unpublished, McMaster University, Hamilton, Ontario, Canada.

Cryer, J. (1986), *Time Series Analysis*, Duxbury Press, Boston, MA.

Dahleh, M., Theodosopoulos, T. and Tsitsiklis, J. (1993), Sample complexity of worst-case identification of FIR linear systems, *in* 'Proceedings of the IEEE Conference on Decision and Control', Vol. 3, pp. 2062–2086.

Davison, A. and Hinkley, D. (1997), *Bootstrap Methods and their Application*, Cambridge University Press, New York, NY.

Dent, W. and Min, A.-S. (1978), 'A monte carlo study of autoregressive integrated moving average processes', *J. Econometrics* **7**, 23–55.

DeVries, D. and van den Hof, P. (1995), 'Quantification of uncertainty in transfer function estimation: a mixed probabilistic-worst-case approach', *Automatica* **31**, 543–557.

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

Draper, N. and Smith, H. (1981), *Applied Regression Analysis, 2nd Ed.*, John Wiley and Sons, New York, NY.

Edgar, T. and Himmelblau, D. (1988), *Optimization of Chemical Processes*, McGraw-Hill, Inc., New York, NY.

Edwards, A. (1972), *Likelihood: An Account of the Statistical Concept of Likelihood and its Application to Scientific Inference*, Cambridge University Press, London, England.

Efron, B. and Hinkley, D. (1978), 'Assessing the accuracy of the maximum likelihood estimator: Observed vs. expected fisher information', *Biometrika* **65**, 457–487.

El-Shaarrawi, A. and Shah, K. (1980), 'Interval estimation in non linear models', *Sankhya* **42**, 227–232.

Eliason, S. (1993), *Maximum Likelihood Estimation, Logic and Practice*, Sage Publications, London, UK.

Fisher, R. (1935), 'The fiducial argument in statistical inference', *The Annals of Eugenics* **6**, 391.

Fisher, R. (1939), 'The sampling distribution of some statistics obtained from non-linear equations', *The Annals of Eugenics* **9**, 238–249.

Fisher, R. (1980), 'A proof of the consistency of maximum likelihood estimators of nonlinear regression models with autocorrelated errors', *Econometrica* **48**, 853–860.

Giarre, L., Milanese, M. and Taragna, M. (1997), 'H_∞ identification and model quality evaluation', *IEEE Trans. Autom. Control* **42**, 188–199.

Goodwin, G. C. and Sin, K. S. (1984), *Adaptive Filtering Prediction and Control*, Prentice-Hall, Englewood Cliffs, N. J.

Goodwin, G., Gevers, M. and Ninness, B. (1992), 'Quantifying the error in estimated transfer functions with application to model order selection', *IEEE Trans. Autom. Control* **37**, 913–927.

Gunnarson, S. (1993), 'On some asymptotic uncertainty bounds in recursive least squares identification', *IEEE Trans. Autom. Control* **38**, 1685–1689.

Gustavsson, I., Ljung, L. and Söderström, T. (1977), 'Identification of processes in closed loop - identifiability and accuracy aspects', *Automatica* **13**, 59–75.

Halperin, M. (1963), 'Confidence interval estimation in non-linear regression', *J. R. Statist. Soc. B* **25**, 330–333.

Halperin, M. and Mantel, N. (1963), 'Interval estimation of non-linear functions', *JASA* **58**, 611–627.

Hamilton, D., Watts, D. and Bates, D. (1982), 'Accounting for intrinsic nonlinearity in nonlinear regression parameter inference regions', *Applied Statistics* **10**, 386–393.

Harris, T. (1977), Theory and application of self-tuning regulators, M.eng. thesis, McMaster University, Hamilton, Ont., Department of Chemical Engineering.

Hartley, H. (1964), 'Exact confidence regions for the parameters in non-linear regression laws', *Biometrika* **51**, 347–353.

Harvey, A. and Phillips, D. (1979), 'Maximum likelihood estimation of regression models with autoregressive-moving average disturbances', *Biometrika* **66**, 49–58.

Hillmer, S. and Tiao, G. (1979), 'Likelihood function of stationary multiple autoregressive moving average models', *JASA* **74**, 652–60.

Isermann, R. (1980), 'Practical aspects of process identification', *Automatica* **16**, 575–587.

Jeffreys, H. (1948), *Theory of Probability, Second Ed.*, Oxford Univ. Press, London, England.

Jun, S. and Tu, D. (1995), *The Jackknife and Bootstrap*, Springer Verlag, New York, NY.

Kang, G. and Rawlings, J. (1998), 'Marginal curvatures for functions of parameters in nonlinear regression', *Stat. Sinica* **8**, 467–476.

Kendall, M. and Stuart, A. (1967), *The Advanced Theory of Statistics, Second Ed.*, Charles Griffin and Comp. Ltd., London, England.

Khorasani, F. (1982), 'Simultaneous confidence bands for nonlinear regression models', *Commun. Statist.-Theory Meth.* **11**, 1241–1253.

Khorasani, F. and Milliken, G. (1979), 'On the exactness of confidence bands about a linear model', *JASA* **74**, 446–448.

Knowles, M., Siegmund, D. and Zhang, H. (1991), 'Confidence regions in semilinear regression', *Biometrika* **78**, 15–31.

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

Lehmann, E. (1959), *Testing Statistical Hypotheses*, John Wiley and Sons, New York, NY.

Lindgren, B. (1976), *Statistical Theory, Third Edition*, MacMillan Publishing Co., Inc., New York, NY.

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

Ljung, G. and Box, G. (1979), 'The likelihood function of stationary autoregressive-moving average models', *Biometrika* **66**, 265–270.

Ljung, L. (1985), 'Asymptotic variance expressions for identified black-box transfer function models', *IEEE Trans. Autom. Control* **30**, 834–844.

Ljung, L. (1987), *System Identification - Theory for the User*, Prentice Hall, Englewood Cliffs, N.J.

Ma, C. (1997), 'On the exact likelihood function of a multivariate autoregressive moving average model', *Biometrika* **84**, 957–964.

Mangasarian, O. (1994), *Nonlinear Programming*, SIAM, Philadelphia.

Mathai, A. and Provost, S. (1992), *Quadratic Forms in Random Variables, Theory and Applications*, Marcel Dekker, New York.

The MathWorks Inc. (1996), *MATLAB v.* 5.1 *User's Guide*, South Natick, MA.

Middleton, R. and Goodwin, G. (1986), 'Improved finite word length characteristics in digital control using delta operators', *IEEE Trans. Autom. Control* **31**, 1015–1021.

Middleton, R. and Goodwin, G. (1990), *Digital Control and Estimation, A Unified Approach*, Prentice-Hall, Inc., New Jersey.

Murdoch, D. (1995), 'Stat 870 course notes', *Queen's University* .

Newbold, P. (1974), 'The exact likelihood function for a mixed autoregressive moving average process', *Biometrika* **61**, 423–426.

Nicholls, D. and Hall, A. (1979), 'The exact likelihood function of multivariate autoregressive-moving average models', *Biometrika* **66**, 259–264.

Ninness, B. and Goodwin, G. (1995a), 'Rapprochement between bounded-error and stochastic estimation theory', *Int. J. Adaptive Control and Signal Processing* **9**, 107–132.

Ninness, B. and Goodwin, G. (1995b), 'Estimation and model quality', *Automatica* **31**, 1771–1797.

Osborn, D. (1977), 'Exact and approximate maximum likelihood estimators for vector moving average processes', *J. R. Statist. Soc. B* **39**, 114–118.

Palmor, Z. and Shinnar, R. (1979), 'Design of sampled data controllers', *Ind. Eng. Chem. Process Des. Dev.* **18**, 8–30.

Palmor, Z. and Shinnar, R. (1981), 'Design of advanced process controllers', *AIChE Journal* **27**, 793–805.

Pankratz, A. (1983), *Forecasting with Univariate Box-Jenkins Models: Concepts and Cases*, John Wiley and Sons, New York, NY.

Phadke, M. and Kedem, G. (1978), 'Computing the exact likelihood function of multivariate moving average models', *Biometrika* **65**, 511–519.

Quinn, S., Bacon, D. and Harris, T. (1999a), 'Assessing the precision of model predictions and other functions of model parameters', *Can. J. Chem. Eng.* **77**, 1–15.

Quinn, S., Bacon, D. and Harris, T. (1999b), 'A note on likelihood intervals and profiling', *submitted for publication* .

Quinn, S., Harris, T. and Bacon, D. (1999c), 'Use of expected profiling for likelihood interval prediction in time series models', *submitted for publication* .

Rao, C. (1962), 'Efficient estimates and optimum inference procedures in large samples', *J. R. Statist. Soc. B* **24**, 46–72.

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

Ravishanker, N. (1994), 'Relative curvature measures of nonlinearity for time series models', *Commun. Statist.-Simula.* **23**, 415–430.

Ravishanker, N., Melnick, E. and Tsai, C.-L. (1990), 'Differential geometry of ARMA models', *J. Time Ser. Anal.* **11**, 259–274.

Reimers, H.-E. (1995), 'Interval forecasting in cointegrated systems', *Statistical Papers* **36**, 349–369.

Reinsel, G. (1980), 'Asymptotic properties of prediction errors for the multivariate autoregressive model using estimated parameters', *J. R. Statist. Soc. B* **42**, 328–333.

Rivera, D., Pollard, J., Sterman, L. and Garcia, C. (1990), An industrial perspective on control-relevant identification, *in* 'Proc. of the 1990 American Control Conf.,', San Diego, CA, USA, pp. 2406–2411.

Ross, G. (1990), *Nonlinear Estimation*, Springer-Verlag, New York, NY.

Roy, T. (1993), An exact confidence interval for the ratio of means using nonlinear regression, *in* 'Eighteenth Annual Conference of the SAS Institute, May, '93', New York, NY. pp. 859-864.

Schoukens, J. and Pintelon, R. (1994), 'Quantifying model errors of identified transfer functions', *IEEE Trans. Autom. Control* **39**, 1733–1737.

Seborg, D. E., Edgar, T. F. and Mellichamp, D. A. (1989), *Process Dynamics and Control*, John Wiley & Sons, New York.

Severini, T. and Staniswalis, J. (1994), 'Quasi-likelihood estimation in semiparametric models', *JASA* **89**, 501–511.

Shirt, R. (1997), Modelling and Identification of Paper Machine Wet End Chemistry, PhD thesis, The University of British Columbia.

Shirt, R., Harris, T. and Bacon, D. (1994), 'Experimental design considerations for dynamic systems', *Ind. Eng. Chem. Res.* **33**, 2656–2667.

Smith, H. and Dubey, S. (1964), 'Some reliability problems in the chemical industry', *Ind. Quality Cont.* **21**, 64–70.

Söderström, T. and Stoica, P. (1989), *System Identification*, Prentice Hall International, London.

Söderström, T., Stoica, P. and Friedlander, B. (1991), 'An indirect prediction error method for system identification', *Automatica* **27**, 183–188.

Sulieman, H. (1998), Empirical and Graphical Methods for Sensitivity Analysis, PhD thesis, Queen's University.

Taniguchi, M. (1986), 'Third order asymptotic properties of maximum likelihood estimates for gaussian arma processes', *J. Multi. Anal.* **18**, 1–31.

Thisted, R. (1988), *Elements of Statistical Computing, Numerical Computation*, Chapman and Hall, London, England.

Tulleken, H. (1990), 'Generalized binary noise test-signal concept for improved identification-experiment design', *Automatica* **26**, 37–49.

van den Hof, P. and Schrama, R. (1995), 'Identification and control - closed-loop issues', *Automatica* **31**, 1751–1770.

van Ewijk, P. and Hoekstra, J. (1994), 'Curvature measures and confidence intervals for the linear logistic model', *Applied Statistics* **43**, 477–487.

Wahlberg, B. and Ljung, L. (1994), 'Hard frequency-domain model error bounds from a least-squares like identification technique', *Automatica* **30**, 391–402.

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

Wei, W. W. S. (1990), *Time Series Analysis: Univariate and Multivariate Methods*, Addison-Wesley, Redwood City, California.

Williams, E. (1962), 'Exact fiducial limits in non-linear estimation', *J. R. Statist. Soc. B* **24**, 125–139.

Yamamoto, T. (1976), 'Asymptotic mean square prediction error for an autoregressive model with estimated coefficients', *Applied Statistics* **25**, 123–127.

Ydstie, B., Kershenbaum, L. and Sargent, R. (1985), 'Theory and application of an extended horizon self-tuning controller', *AIChE Journal* **31**, 1771–1780.

Young, D., Zerbe, G. and Hay, W. (1997), 'Fieller's theorem, scheffe simultaneous confidence intervals, and ratios of parameters of linear and nonlinear mixed-effects models', *Biometrics* **53**, 838–847.

Zarrop, M. (1979), *Optimal Experiment Design for Dynamic System Identification*, Springer-Verlag, Berlin.

Zhou, T. and Kimura, H. (1992), 'Simultaneous identification of nominal model, parametric uncertainty and unstructured uncertainty for robust control', *IEEE Trans. Autom. Control* **37**, 900–912.

Zhu, Y. (1989), 'Estimation of transfer functions: Asymptotic theory and a bound on model uncertainty', *Int. J. Control* **49**, 2241–2258.

Appendix A

Appendix Outlining Computational Issues

The purpose of this appendix is to provide any details about the computational methods which were omitted from the manuscripts or which are so important as to be worth repeating here. All computation and visualization was done using MATLAB^{TM} version 4.2c or 5.1.

A.1 Generalized Profiling

Two different algorithms for generalized profiling are given in Figures A.1 and A.2. The first algorithm (A.1) is based on an optimization approach to generalized profiling, and the second (A.2) is based on reparameterization. MATLAB's version of the Simplex Algorithm of Nelder and Mead (Edgar and Himmelblau, 1988) was used to solve all unconstrained optimization problems. All constrained optimization problems were solved using a Sequential Quadratic Programming routine (Edgar and Himmelblau, 1988). The value of Δ (see Figures A.1 and A.2) was, in most cases, set to $\pm\frac{se(\hat{g})}{6}$. Bates and Watts (1988) suggested that the value of Δ be adjusted at each iteration based on the slope of the profile t curve in the previous iteration. We have found that performance of the algorithms is best if Δ is left constant unless convergence problems are encountered, at which point we reassign $\Delta = sign(\Delta)(max[\Delta/2, se(\hat{g})/24])$.

A cubic spline was fitted to calculated values of $g(\boldsymbol{\theta})$ versus τ for the purpose of obtaining the limits of the likelihood interval. This proved satisfactory, although other methods of interpolation would likely also perform well in most cases.

In this work, the variance of the random error was computed as:

$$s^2 = \frac{S(\boldsymbol{\theta})}{n-p} \tag{A.1}$$

where $S(\boldsymbol{\theta})$ is the sum of squared residuals; however, any other available estimate of s^2 could be used, with the appropriate change in the number f degrees of freedom. For dynamic models (i.e., time series models and SISO transfer function models) the

1. Using a nonlinear optimization package, find the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$.

2. Compute the MLE of $g(\boldsymbol{\theta})$, and define $\hat{g} = g(\hat{\boldsymbol{\theta}})$.

3. Compute an estimate of the variance of the random error (i.e., compute $s^2 = \frac{S(\hat{\boldsymbol{\theta}})}{n-p}$ with $n - p$ degrees of freedom).

4. Compute $Cov(\hat{\boldsymbol{\theta}})$.

5. Compute $se(\hat{g}) = \sqrt{s^2 \left.\frac{dg}{d\boldsymbol{\theta}}^T\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{V}^T\boldsymbol{V})^{-1} \left.\frac{dg}{d\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}}$.

6. Set the index i to 1, and let $g_{old} = \hat{g}$.

7. Move the value of $g(\boldsymbol{\theta})$ away from $\hat{g}$ by a small amount Δ (i.e., $g_i = g_{old} + \Delta$).

8. Use a constrained nonlinear optimization package to solve the constrained optimization problem: maximize $L(\boldsymbol{\theta})$ subject to $g(\boldsymbol{\theta}) = g_i$. The location of the constrained optimum is $\bar{\boldsymbol{\theta}}$.

9. Compute

$$\tau_i = sign(g_i - \hat{g})\sqrt{-2ln\left(\frac{L(\bar{\boldsymbol{\theta}})}{L(\hat{\boldsymbol{\theta}})}\right)}$$

$$\delta_i = \frac{g_i - \hat{g}}{se(\hat{g})}$$

10. Is $|\tau_i| \geq t(n-p, \alpha/2)$? If yes, continue. If no, set $g_{old} = g_i$, set $i = i + 1$, and return to Step 7.

11. Is Δ negative? If yes, continue. If no, set $g_{old} = \hat{g}$, set $i = i + 1$, let $\Delta = -\Delta$ and return to Step 7.

12. Fit a smooth curve through the values of τ_i and use this curve to find the values of $g(\boldsymbol{\theta})$ at $\tau = \pm t(n-p, \alpha/2)$. These are the limits of the likelihood interval for $g(\boldsymbol{\theta})$.

13. Compute the limits of the linearization confidence interval for $g(\boldsymbol{\theta})$ using

$$CI = \hat{g} \pm se(\hat{g})t(n-p, \alpha/2)$$

14. Construct the profile t plot by plotting, on one figure, τ_i versus g_i, and δ_i versus g_i.

Figure A.1: A step-by-step algorithm for the optimization approach to generalized profiling.

1. Using a nonlinear optimization package, find the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$.

2. Compute the MLE of $g(\boldsymbol{\theta})$, and define $\hat{g} = g(\hat{\boldsymbol{\theta}})$.

3. Compute an estimate of the variance of the random error (i.e., compute $s^2 = \frac{S(\hat{\boldsymbol{\theta}})}{n-p}$ with $n-p$ degrees of freedom).

4. Compute $Cov(\hat{\boldsymbol{\theta}})$.

5. Compute $se(\hat{g}) = \sqrt{s^2 \left.\frac{dg}{d\boldsymbol{\theta}}^T\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\boldsymbol{V}^T\boldsymbol{V})^{-1} \left.\frac{dg}{d\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}}$.

6. Reparameterize the model such that the first new parameter is $\phi_1 = g(\boldsymbol{\theta})$.

7. Set the index i to 1, and let $g_{old} = \hat{g}$

8. Move the value of ϕ_1 away from $\hat{g}$ by a small amount Δ (i.e., $g_i = g_{old} + \Delta$) .

9. Create a new vector of $p-1$ parameters which does not include ϕ_1. Let the new vector be $\boldsymbol{\phi}_{reduced}$, where the parameters in $\boldsymbol{\phi}_{reduced}$ are functions of $\boldsymbol{\theta}$.

10. Use an unconstrained optimization package to locate the conditional maximum likelihood estimate of $\boldsymbol{\phi}_{reduced}$. This estimate is $\hat{\boldsymbol{\phi}}_{reduced}$.

11. In p-dimensional space, the location of the conditional maximum likelihood estimate is $\tilde{\boldsymbol{\phi}} = [g_i, \hat{\boldsymbol{\phi}}^T_{reduced}]$.

12. Compute

$$\tau_i = sign(g_i - \tilde{\phi}_1)\sqrt{-2ln\left(\frac{L(\tilde{\boldsymbol{\phi}})}{L(\hat{\boldsymbol{\phi}})}\right)}$$

$$\delta_i = \frac{g_i - \tilde{\phi}_1}{se(\tilde{\phi}_1)}$$

13. Is $|\tau_i| \geq t(n-p, \alpha/2)$? If yes, continue. If no, set $g_{old} = g_i$, set $i = i+1$, and return to Step 8.

14. Is Δ negative? If yes, continue. If no, set $g_{old} = \hat{g}$, set $i = i+1$, let $\Delta = -\Delta$ and return to Step 8.

15. Fit a smooth curve through the values of τ_i and use this curve to find the values of $g(\boldsymbol{\theta})$ at $\tau = \pm t(n-p, \alpha/2)$. These are the limits of the likelihood interval for $g(\boldsymbol{\theta})$.

16. Compute the limits of the linearization confidence interval for $g(\boldsymbol{\theta})$ using

$$CI = \hat{g} \pm se(\hat{g})t(n-p, \alpha/2)$$

17. Construct the profile t plot by plotting, on one figure, τ_i versus g_i, and δ_i versus g_i.

Figure A.2: A step-by-step algorithm for the reparameterization approach to generalized profiling.

variance of the white noise innovation was estimated by

$$s^2 = \frac{S^*(\boldsymbol{\theta})}{n-p} \tag{A.2}$$

where $S^*(\boldsymbol{\theta})$ is the modified sum of squares (Ansley, 1979). It is computed as follows. Let $\boldsymbol{\Omega}$ be the variance-covariance matrix for z_t. The Cholesky decomposition of $\boldsymbol{\Omega}$ is

$$\boldsymbol{\Omega} = \boldsymbol{L}\boldsymbol{L}^T \tag{A.3}$$

where $\boldsymbol{L}$ is a lower triangular band matrix. Then,

$$\tilde{\boldsymbol{e}} = |\boldsymbol{L}|^{1/n}\boldsymbol{L}^{-1}\boldsymbol{z} \tag{A.4}$$

where $\boldsymbol{z}$ is an $n \times 1$ vector of values of z_t. Then, the modified sum of squared residuals is

$$S^*(\boldsymbol{\theta}) = \sum_{i=1}^{n} \tilde{e}_i^2 \tag{A.5}$$

For all models, the covariance matrix for $\hat{\boldsymbol{\theta}}$ was calculated using:

$$Cov(\hat{\boldsymbol{\theta}}) = (\boldsymbol{V}^T\boldsymbol{V})^{-1} \tag{A.6}$$

where $\boldsymbol{V} = \left.\frac{\partial f(\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$.

Donaldson and Schnabel (1987) identified three different ways by which the covariance matrix of the estimated parameters of regression models can be approximated:

$$Cov(\hat{\boldsymbol{\theta}}) = (\boldsymbol{V}^T\boldsymbol{V})^{-1} \tag{A.7}$$

$$Cov(\hat{\boldsymbol{\theta}}) = (\boldsymbol{H})^{-1} \tag{A.8}$$

$$Cov(\hat{\boldsymbol{\theta}}) = (\boldsymbol{H})^{-1}(\boldsymbol{V}^T\boldsymbol{V})(\boldsymbol{H})^{-1} \tag{A.9}$$

where $\boldsymbol{H}$ is the Hessian matrix of $S(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, with elements

$$\boldsymbol{H}_{ij} = \left.\frac{\partial^2 S(\boldsymbol{\theta})}{\partial\theta_i\,\partial\theta_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{A.10}$$

For use with a more general class of models, $S(\boldsymbol{\theta})$ can be replaced by $\ln L(\boldsymbol{\theta})$. Donaldson and Schnabel found little difference between these three estimates of the covariance matrix, and so we have chosen to use the simplest and most common expression (A.7).

For dynamic models, the covariance matrix of $\hat{\boldsymbol{\theta}}$ can be based on the expected covariance matrix which is a function of the model and any exogenous inputs, and does not depend on the data. This is the method we use in expected profiling where we are not considering a specific realization of data. The *expected* Fisher Information matrix

$\mathcal{I}_{\hat{\boldsymbol{\theta}}}$ is used to compute the expected variance-covariance matrix of the parameter estimates, where

$$\left\{\mathcal{I}_{\hat{\boldsymbol{\theta}}}\right\}_{ij} = E\left\{-\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right\}\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \tag{A.11}$$

(A.12)

and $\boldsymbol{\theta}_0$ represents the true values of the parameters. In practice, the variance-covariance matrix for $\hat{\boldsymbol{\theta}}$ can be based on the Cramer-Rao lower bound (Ljung, 1987):

$$cov(\hat{\boldsymbol{\theta}}) \geq \sigma_a^2 [\mathcal{I}]^{-1} \tag{A.13}$$

$$\geq \sigma_a^2 \left[\sum_{t=1}^{n} E\left\{\psi(t,\theta_0)\psi^T(t,\theta_0)\right\}\right]^{-1} \tag{A.14}$$

where

$$\psi(t,\theta_0) = \frac{\partial}{\partial\boldsymbol{\theta}} a(t,\theta) \tag{A.15}$$

In the case where a realization of data is available, we estimate the Fisher Information matrix on the basis of the data. That is, we compute the *observed* Fisher information matrix $I_{\hat{\boldsymbol{\theta}}}$ using

$$\left\{I_{\hat{\boldsymbol{\theta}}}\right\}_{ij} = \left[-\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right]\Bigg|_{\hat{\boldsymbol{\theta}}} \tag{A.16}$$

Efron and Hinkley (1978) provided evidence that $I_{\hat{\boldsymbol{\theta}}}$ should be preferred over $\mathcal{I}_{\hat{\boldsymbol{\theta}}}$ for computing $cov(\hat{\boldsymbol{\theta}})$.

Other methods for computing the covariance matrices for parameter estimates in various models have been proposed. Many of these alternative methods are resampling methods (see, for example, Spall, 1998), and are computationally intensive. These were not used in the work being presented here.

In this work, we computed the observed Fisher Information matrix $I_{\hat{\boldsymbol{\theta}}}$ in all cases where data is involved. However, we have found that for time series models and transfer function models, the variance approximations for individual parameter estimates based on $I_{\hat{\boldsymbol{\theta}}}$ may be poor if the vector of parameters $\hat{\boldsymbol{\theta}}$ is close to a stability or invertibility boundary. The error in the variance estimate manifests itself on a profile t plot as a reference line which is not tangent to the profile at $\hat{\boldsymbol{\theta}}$. It has been proven (see Chapter 4) that the linear approximation reference line should be tangent to the profile t plot at $\hat{\boldsymbol{\theta}}$; therefore, we have adjusted the linear approximation reference lines appropriately so that they are tangent to their associated profiles.

For an ARMA(p,q) model $\phi(q^{-1})y_t = \theta(q^{-1})a_t$, the expressions for the derivatives of a_t with respect to the parameters are (Ravishanker, 1994):

$$\frac{\partial a_t}{\partial \phi_k} = -\frac{1}{\phi(q^{-1})} a_{t-k} = -\frac{1}{\theta(q^{-1})} y_{t-k} \tag{A.17}$$

$$\frac{\partial a_t}{\partial \theta_l} = \frac{1}{\theta(q^{-1})} a_{t-l} = \frac{\phi(q^{-1})}{\theta^2(q^{-1})} y_{t-l} \tag{A.18}$$

Let

$$\upsilon(u) = cov\left(\frac{\partial a_t}{\partial \phi_k}, \frac{\partial a_{t+u}}{\partial \phi_k}\right) = cov\left(\frac{\partial a_t}{\partial \phi_k}, \frac{\partial a_t}{\partial \phi_{k+u}}\right) \tag{A.19}$$

$$\nu(u) = cov\left(\frac{\partial a_t}{\partial \theta_l}, \frac{\partial a_{t+u}}{\partial \theta_l}\right) = cov\left(\frac{\partial a_t}{\partial \theta_l}, \frac{\partial a_t}{\partial \theta_{l+u}}\right) \tag{A.20}$$

(Åström, 1980) and

$$\varrho_{\phi\theta}(u) = cov\left(\frac{\partial a_t}{\partial \phi_k}, \frac{\partial a_{t+u}}{\partial \theta_l}\right) \tag{A.21}$$

Note that

$$\varrho_{\phi\theta}(u) \neq \varrho_{\theta\phi}(u) \tag{A.22}$$

but

$$\varrho_{\phi\theta}(u) = \varrho_{\theta\phi}(-u) \tag{A.23}$$

Then, in order to calculate the lower bound for $Cov(\hat{\boldsymbol{\theta}})$ based on (A.14) so as to obtain a value for $se(g)$, we develop the expression:

$$\begin{aligned} \Upsilon &= E\{\mathcal{I}\} \\ &= E\left\{\boldsymbol{\psi}\boldsymbol{\psi}^{\boldsymbol{T}}\right\} \\ &= \begin{bmatrix} \upsilon(0) & \cdots & \upsilon(p-1) & \varrho_{\phi\theta}(0) & \cdots & \varrho_{\phi\theta}(q-1) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \upsilon(p-1) & \cdots & \upsilon(0) & \varrho_{\phi\theta}(1-p) & \cdots & \varrho_{\phi\theta}(0) \\ \varrho_{\phi\theta}(0) & \cdots & \varrho_{\phi\theta}(1-p) & \nu(0) & \cdots & \nu(q-1) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \varrho_{\phi\theta}(q-1) & \cdots & \varrho_{\phi\theta}(0) & \nu(q-1) & \cdots & \nu(0) \end{bmatrix} \end{aligned} \tag{A.24}$$

We compute the elements of this covariance matrix by first computing the impulse responses of the $\frac{\partial a_t}{\partial \phi_k}$ and $\frac{\partial a_t}{\partial \theta_l}$. For example, to compute the covariance between:

$$w_t = \frac{\partial a_t}{\partial \phi_k} = \frac{-1}{\phi(q^{-1})} a_{t-k} \tag{A.25}$$

and

$$x_t = \frac{1}{\theta(q^{-1})} a_{t-l} \tag{A.26}$$

first compute the impulse response of each system. That is:

$$\begin{aligned} w_t &= (1+\zeta_1 B+\zeta_2 B^2+\ldots)a_{t-k} \\ x_t &= (1+\xi_1 B+\xi_2 B^2+\ldots)a_{t-l} \end{aligned} \tag{A.27}$$

then,

$$\varrho_{xw}(u) = E\{x_t w_{t+u}\} = \sigma^2 \sum_{m=0}^{\infty} \zeta_{m+u}\xi_m \tag{A.28}$$

$$\varrho_x(u) = E\{x_t x_{t+u}\} = \sigma^2 \sum_{m=0}^{\infty} \xi_{m+u}\xi_m \tag{A.29}$$

Although the expressions (A.29) and (A.28) are not exact unless the infinite sum is computed, in practice, the error incurred by truncating the sum at an appropriately high lag is negligible unless one or more of the series is virtually unstable. In such cases, the variances should be computed using an exact analytical method. Åström (1970) showed that the variance at lag 0 can be calculated by

$$\varrho_x(0) = \frac{\sigma^2}{2\pi}\oint \frac{P(z)P(z^{-1})}{Q(z)Q(z^{-1})}\frac{dz}{z} \tag{A.30}$$

$$= \frac{\sigma^2}{2\pi}\int_{-\pi}^{\pi}\left|\frac{P(e^{i\omega})}{Q(e^{i\omega})}\right| d\omega \tag{A.31}$$

where $P(z)$ is the moving average polynomial and $Q(z)$ is the autoregressive polynomial. Crowe (1976) (as reported in Harris, 1977) extended this method to autocovariances at lag k by developing the expression

$$\varrho_x(k) = \frac{\sigma^2}{4\pi i}\left(\oint \frac{P_1(z)P_1(z^{-1})}{Q(z)Q(z^{-1})}\frac{dz}{z} - \oint \frac{P_2(z)P_2(z^{-1})}{Q(z)Q(z^{-1})}\frac{dz}{z}\right) \tag{A.32}$$

where z is the z-transform operator,

$$P_1(z) = P(z)\left(\cos\left(\frac{5}{4}\pi\right)z^k + \cos\left(\frac{3}{4}\pi\right)\right) \tag{A.33}$$

and

$$P_2(z) = P(z)\left(\sin\left(\frac{5}{4}\pi\right)z^k + \sin\left(\frac{3}{4}\pi\right)\right) \tag{A.34}$$

This can also be written as

$$\varrho_x(k) = \frac{\sigma^2}{2\pi}\oint \frac{P(z)P(z^{-1})}{Q(z)Q(z^{-1})}\frac{dz}{z} \tag{A.35}$$

$$= \frac{\sigma^2}{2\pi}\int_{-\pi}^{\pi}\left|\frac{P(e^{i\omega})}{Q(e^{i\omega})}\right| e^{i\omega k} d\omega \tag{A.36}$$

Ljung (1987) presents an efficient algorithm for computing A.36 based on an algorithm derived by Åström (1970). Ljung's algorithm is as follows. Given the model

$$Q(z)y_t = P(z)a_t \tag{A.37}$$

where $Q(z) = q_0 z^r + q_1 z^{r-1} + \ldots + q_r$ and $P(z) = p_0 z^r + p_1 z^{r-1} + \ldots + p_r$, let $p_i^r = p_i$ and $q_i^r = q_i$, and define p_i^k and q_i^k recursively by

$$p_i^{r-k} = \frac{q_0^{r-k+1} p_i^{r-k+1} - p_{r-k+1}^{r-k+1} q_{r-k+1-i}^{r-k+1}}{q_0^{r-k+1}} \tag{A.38}$$

$$q_i^{r-k} = \frac{q_0^{r-k+1} q_i^{r-k+1} - q_{r-k+1}^{r-k+1} q_{r-k+1-i}^{r-k+1}}{q_0^{r-k+1}} \tag{A.39}$$

then

$$\varrho_0 = \frac{1}{q_0} \sum_{k=0}^{r} \frac{(p_k^k)^2}{q_0^k} \tag{A.40}$$

Cross covariances at lag k are not readily computed by these methods (Harris, 1977).

A.2 Reparameterization

In some cases, it is straightforward to find an algebraic solution to the reparameterization when using the reparameterization algorithm given in Figure A.2. However, in other cases, it is difficult, or even impossible, to find an explicit expression for the model in terms of the vector of new parameters $\boldsymbol{\phi}$. In such cases, the reparameterization may be carried out numerically using a nonlinear equation solver, or a constrained optimization package. In this work, all reparameterization was done using MATLAB's equation solver.

Even when using the optimization approach to generalized profiling, a reparameterization of the model was sometimes used to implicitly constrain parameters to lie within a feasible region. For example, the parameters of the Michaelis-Menton model

$$y = \frac{\theta_1 x}{\theta_2 + x} + \epsilon \tag{A.41}$$

must all be positive. By reparameterizing the model such that

$$\theta_1 = \phi_1^2 \tag{A.42}$$

$$\theta_2 = \phi_2^2 \tag{A.43}$$

and performing all optimizations in terms of ϕ_1 and ϕ_2, we can implicitly ensure that both θ_1 and θ_2 remain positive while still using an unconstrained optimization algorithm.

Similarly, when profiling parameters and functions of parameters of time series and transfer function models, the partial autocorrelation function (PACF) transformation was used to ensure that all parameter vectors remained within the stability/invertibility region. For the general transfer function model

$$A(q^{-1})yt = \frac{B(q^{-1})}{F(q^{-1})}u_{t-d} + \frac{C(q^{-1})}{D(q^{-1})}e_t \tag{A.44}$$

we reparameterize each of the polynomials $A(q^{-1})$, $C(q^{-1})$, $D(q^{-1})$ and $F(q^{-1})$ individually using the definition of the PACF for autoregressive (AR) processes. The PACF function is defined by

$$\phi_{kk} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & \rho_k \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_{k-2} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & \rho_1 \end{vmatrix}} \tag{A.45}$$

where ϕ_{kk} is the k^{th} partial autocorrelation and ρ_k is the k^{th} autocorrelation (Wei, 1990). Cryer (1986) provides the following recursive algorithm for computing ϕ_{kk}

$$\phi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1} \phi_{k-1,j}\rho_{k-j}}{1 - \sum_{j=1}^{k} \phi_{k-1,j}\rho_j} \tag{A.46}$$

where $\phi_{k,j} = \phi_{k-1,j} - \phi_{kk}\phi_{k-1,k-j}$.

That is, we treat each individual polynomial as if it were the polynomial defining an AR process, and then find the values of the PACF on this basis. The parameter vector consists of all p PACF values computed in this way. The advantage of this transformation is that the individual stability/invertibility limits for the new parameters are ± 1, and the p-dimensional stability region is a hypercube with sides located at ± 1 (Åström and Wittenmark, 1990). This property allows the new parameters to be easily constrained to remain within the stability/invertibility region since most constrained optimization packages accept bounds on individual parameters as inputs to the routine. The PACF transformation is also used in Chapter 6 as the basis for a measure of nonlinearity for ARMA models. It has been our experience that explicitly constraining the parameters of dynamic models to remain within the stability/invertibility boundary significantly improves the chance of converging to an optimum, and also improves the rate at which that convergence is achieved.

Two reparameterizations were used to constrain the parameters. The first is the PACF transformation which has already been discussed. In cases where the solution

to an otherwise unconstrained optimization problem was required, the parameters were transformed as follows.

Let θ_i represent any one parameter of an ARMA(p,q) model and $\phi_{i,max}$ be the upper stability bound for that parameter, i.e.

$$-\phi_{i,max} \leq \theta_i \leq \phi_{i,max} \tag{A.47}$$

Then,

$$-1 \leq \frac{\theta_i}{\phi_{i,max}} \leq 1 \tag{A.48}$$

and,

$$0 \leq \frac{\theta_i}{2\phi_{i,max}} + 0.5 \leq 1 \tag{A.49}$$

Let $P_N(x \leq u)$ be the probability that a Normally distributed random variable x with mean zero and unit variance is less than u. u may range over all real numbers but the range of $P_N(x \leq u)$ is $[0, 1]$.

Let the set of new parameters following reparameterization of the model be $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ such that

$$\gamma_i = u \tag{A.50}$$

where u satisfies $P_N(x \leq u) = \frac{\theta_i}{2\phi_{i,max}} + 0.5$. When profiling an individual parameter of a model, the optimization problem solved during each iteration of the profiling algorithm is carried out in $p-1$ dimensions (see Figure A.2). Therefore, the optimization for $\boldsymbol{\gamma}$ is carried out over the set of real numbers in $p-1$ directions and the solution is inversely transformed to obtain the results in terms of the original parameters $\boldsymbol{\theta}$. That is, the profiling is done in terms of γ_i and then for each γ_i, θ_i is found as follows:

$$\theta_i = 2\phi_{i,max}(P_N(x \leq \gamma_i) - 0.5) \tag{A.51}$$

Parameters so transformed do not have individual upper or lower limits. However, the stability/invertibility region is still a subset of $\Re$, and therefore convergence problems may arise due to the vector of parameters wandering outside the acceptable region. For our calculations, a penalty function was always used in conjunction with the transformation given in (A.50).

A.3 Stationarity, Stability and Invertibility

The likelihood functions for parameters in time series and transfer function models approach negative infinity asymptotically as a stability boundary is approached, and therefore, the nature of the likelihood function implicitly constrains the solutions of modeling and inference problems to remain within the stable (stationary) region.

However, the moving average parameters are not implicitly constrained by the likelihood function. Therefore, all optimization routines were implemented such that the moving average parameters were explicitly constrained to lie within the invertibility boundaries. In some cases this was achieved through reparameterization as discussed above. In other cases, a penalty function approach was taken. In these cases, the likelihood function was forced to take on very large negative values whenever the parameter vector moved outside of the stability/invertibility region.

A.4 Transfer Function Models

There have been many approaches used to estimate transfer functions from data. Some of the more common classes of methods include: prediction error methods (Söderström et al., 1991; Ljung, 1987; Åström, 1980), frequency domain methods (Ljung, 1987), and maximum likelihood methods (Åström, 1980; Ljung, 1987). Also, there is a large body of literature devoted to exploring the special issues involved in estimating process models from closed-loop data, and establishing explicit algorithms for this estimation situation (van den Hof and Schrama, 1995; Gustavsson et al., 1977).

Recently, attention has been focused on control-relevant identification (van den Hof and Schrama, 1995; Ljung, 1987; Rivera et al., 1976). With this approach the criteria and cost function which are optimized to obtain a "best" model are based on control-relevant indices.

The focus of this thesis is generalized profiling – a likelihood ratio approach to estimating uncertainty in functions of parameters. Therefore, model identification is focused on the method of maximum likelihood and approximate likelihood methods. Because alternative estimation criteria (any criterion other than maximum likelihood) are widely used in control, the use of profiling in this context is considered briefly in Chapter 7.

Ansley and Newbold (1980) analyzed, via simulation, the properties of three commonly used estimation methods for fitting dynamic models to data: maximum likelihood, exact least squares, and a conditional likelihood method (which is described later in this chapter). Although no one method performed best in all cases, the method of maximum likelihood resulted in estimates with low bias and models with high predictive power.

Full maximum likelihood estimates are often based on Kalman filtering (Åström, 1980; Gustavsson et al., 1977). However, we have used a maximum likelihood approach described in the time series literature (Ansley, 1979). Ansley's algorithm is based on a clever transformation of an ARMA-type model. The procedure for using Ansley's algorithm in the context of transfer function models is given in Figure A.3. The description given in Figure A.3 was used so as to be consistent with the work done on expected profiling (see Chapter 5), which was also based on this method. However, the Ansley's approach is very computationally intensive.

The approximate likelihood method is based on an approximation to the likelihood function. If starting values for the series a_t and y_t were known, then, for a known

1. Write the transfer function model in the form

$$y_t = \frac{B(q^{-1})}{A(q^{-1})F(q^{-1})}q^{-k}u_t + \frac{C(q^{-1})}{A(q^{-1})D(q^{-1})}a_t \tag{A.52}$$

2. Assume that all u_t are known, even for $t < 0$. Alternatively, assume that the process was operating at steady state prior to the experiment, i.e., $u_t = 0$, $t < 0$, where u_t is expressed in terms of deviations from set point.

3. Then, using initial estimates of the parameters, compute the filtered series

$$u_f = \frac{B(q^{-1})}{A(q^{-1})F(q^{-1})}q^{-k}u_t \tag{A.53}$$

4. Compute

$$w_t = y_t - u_f \tag{A.54}$$

5. Now write the model as a time series

$$\phi(q^{-1})w_t = \theta(q^{-1})a_t \tag{A.55}$$

where $\phi(q^{-1}) = A(q^{-1})D(q^{-1})$ and $\theta(q^{-1}) = C(q^{-1})$.

6. Ansley's transformation is

$$z_t = \begin{cases} w_t, & t = 1, \ldots, m \\ \phi(q^{-1})w_t, & t = m+1, \ldots, n \end{cases} \tag{A.56}$$

where $m = max(na \times nd, nc)$, and na, nb, nc and nd are the degrees of the polynomials $A(q^{-1})$, $B(q^{-1})$, $C(q^{-1})$ and $D(q^{-1})$, respectively. Let:

$$v_t = \theta(q^{-1})a_t = \phi(q^{-1})w_t \tag{A.57}$$

The series v_t is autocorrelated only up to lag q. Then, the covariance matrix for z_t has a maximum bandwidth of m for the first m rows and a bandwidth of nc thereafter (Ansley, 1979). See Ansley (1979) for an efficient way to compute the likelihood based on the transformed series z_t.

Figure A.3: A step-by-step algorithm for maximum likelihood estimation of a SISO transfer function model (modified from Ansley, 1979).

variance σ_1^2, the likelihood function of the model parameters $\boldsymbol{\theta}$ could be written as:

$$L(\boldsymbol{\theta}) = (2\pi\sigma_a^2)^{-n/2} \exp\left(\frac{-\boldsymbol{a}^T\boldsymbol{a}}{2\sigma_a^2}\right) \tag{A.58}$$

and the maximum likelihood estimates would be equal to the least squares estimates. Using this result, conditional maximum likelihood estimates can be calculated by setting the starting values of a_t and y_t to their expected values, and then choosing $\boldsymbol{\theta}$ such that it minimizes $\boldsymbol{a}^T\boldsymbol{a}$ (Al'tshuler, 1983). Box and Jenkins (1976) argue that the determinant of the covariance matrix is dominated by the exponential term and therefore the determinant may be disregarded when n is large. However, this approach may lead to inferior estimators (Dent and Min, 1978). Box and Jenkins (1976) also proposed an approximation based on backforecasting. However, if the parameters of the model are close to a stability/invertibility boundary, then this method may be poor or require numerous iterations.

We compute approximate MLE by maximizing (A.58) with respect to the parameters *and* the starting conditions. Although the resulting estimates are not the true maximum likelihood estimates (Thisted, 1988), they are excellent approximations.

For both the full and approximate likelihood algorithms, convergence problems may be encountered if the model is overparameterized. A model having common roots in its polynomials in q^{-1} will have a covariance matrix $\boldsymbol{\Omega}$ which is singular. When trying to estimate a model which is overparameterized, or which has a numerator and denominator with similar roots, convergence problems may occur because $\boldsymbol{\Omega}$ will be ill-conditioned.

A.5 Expected Profiling

A stepwise procedure for expected profiling is given in Figure A.4. The expression for $E\{\tau^2\}$ is

$$E\{\tau^2\} = tr\left(\tilde{\boldsymbol{\Omega}}_n{}^{-1}\boldsymbol{\Omega}_n^*\right) - n + \ln\left(\frac{\left|\tilde{\boldsymbol{\Omega}}_n\right|}{|\boldsymbol{\Omega}_n^*|}\right) \tag{A.60}$$

Note that to compute the first term of this expression, the two $n \times n$ covariance matrices $\tilde{\boldsymbol{\Omega}}_n$ and $\boldsymbol{\Omega}_n^*$ must be found. Then, the inverse of $\tilde{\boldsymbol{\Omega}}_n$ must be taken, and the product of $\tilde{\boldsymbol{\Omega}}_n{}^{-1}$ and $\boldsymbol{\Omega}_n^*$ computed. For large n, this may involve a prohibitively large number of calculations. Therefore, there is a need for an efficient algorithm by which to compute $tr(\tilde{\boldsymbol{\Omega}}_n{}^{-1}\boldsymbol{\Omega}_n^*)$. The models and the values of the parameters define the values of the elements in the matrices $\tilde{\boldsymbol{\Omega}}_n$ and $\boldsymbol{\Omega}_n^*$. These values are *not* based on measured data, but are computed on the basis of the expected covariance structure defined by the model (see Equation A.16).

Lam and Watts (1991) based their profiling calculations on the expression developed by Ansley (1979) for the exact likelihood function of an ARMA model. However, many others (e.g. Newbold, 1974;, Ali, 1977; and Ljung and Box, 1979) have proposed

1. Define the form of the model and chose values for its parameters. Set $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$, where $\boldsymbol{\theta}^*$ is the vector of "true" values of the parameters.

2. Choose a value for n, the proposed length of the time series.

3. Compute the expected variance-covariance matrix $Cov(\hat{\boldsymbol{\theta}})$ for $\boldsymbol{\theta}$ based on (A.14). This is defined by the model and the values of its parameters.

4. Define $\hat{g} = g(\hat{\boldsymbol{\theta}})$.

5. Compute $se(\hat{g}) = \sqrt{s^2 \left.\frac{dg}{d\boldsymbol{\theta}}^T\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (Cov(\hat{\boldsymbol{\theta}}))^{-1} \left.\frac{dg}{d\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}}$.

6. Set the index i to 1, and let $g_{old} = \hat{g}$.

7. Move the value of $g(\boldsymbol{\theta})$ away from $\hat{g}$ by a small amount Δ (i.e., $g_i = g_{old} + \Delta$).

8. Use a constrained nonlinear optimization package to solve the constrained optimization problem: maximize

$$-ln\left|\tilde{\boldsymbol{\Omega}}_n\right| - tr\left(\tilde{\boldsymbol{\Omega}}_n{}^{-1}\boldsymbol{\Omega}_n^*\right) \tag{A.59}$$

   subject to $g(\boldsymbol{\theta}) = g_i$. The location of the constrained optimum is $\tilde{\boldsymbol{\theta}}$.

9. Compute

$$\tau_i = sign(g_i - \hat{g})\sqrt{tr\left(\tilde{\boldsymbol{\Omega}}_n{}^{-1}\boldsymbol{\Omega}_n^*\right) - n + ln\left(\frac{\left|\tilde{\boldsymbol{\Omega}}_n\right|}{\left|\boldsymbol{\Omega}_n^*\right|}\right)}$$

$$\delta_i = \frac{g_i - \hat{g}}{se(\hat{g})}$$

10. Is $|\tau_i| \geq t(n-p, \alpha/2)$? If yes, continue. If no, set $g_{old} = g_i$, set $i = i + 1$, and return to Step 7.

11. Is Δ negative? If yes, continue. If no, set $g_{old} = \hat{g}$, set $i = i + 1$, let $\Delta = -\Delta$ and return to Step 7.

12. Fit a smooth curve through the values τ_i and use this curve to find the values of $g(\boldsymbol{\theta})$ at $\tau = \pm t(n-p, \alpha/2)$. These are the limits of the likelihood interval LI_n for $g(\boldsymbol{\theta})$.

13. Compute the limits of the linearization confidence interval for $g(\boldsymbol{\theta})$ using

$$CI_n = \hat{g} \pm se(\hat{g})t(n-p, \alpha/2)$$

14. Construct the profile t plot by plotting, on one figure, τ_i versus g_i, and δ_i versus g_i.

15. Repeat from Step 2 for a different value of n. If all values of n of interest have been profiled, proceed to Step 16.

16. Construct the n-plot by plotting CI_n versus n and LI_n versus n.

Figure A.4: A step-by-step algorithm for expected profiling.

expressions and algorithms for computing the exact likelihood. Harvey and Phillips (1979), and Åström (1980), among others have developed algorithms based on the Kalman filter. Expressions for the exact likelihood function for vector ARMA(p,q) processes have been developed by Osborn (1977), Phadke and Kedem (1978), Hillmer and Tiao (1979), and Nicholls and Hall (1979). These, of course, can also be used for the special case of univariate ARMA models.

To be consistent with the previous work, we have based our calculations on the following transformation proposed by Ansley (1979). The expressions for $\tilde{\mathbf{\Omega}}_n$ and $\mathbf{\Omega}^*_n$, which are obtained from the transformation, are developed below.

Ansley's algorithm is based on transforming an ARMA(p,q) time series y_t as follows:

$$z_t = \begin{cases} y_t, & t = 1, \ldots, m \\ \phi(B)y_t, & t = m+1, \ldots, n \end{cases} \tag{A.61}$$

where $m = max(p, q)$. Let:

$$v_t = \theta(B)a_t = \phi(B)y_t \tag{A.62}$$

The series v_t is autocorrelated only up to lag q. Then, the covariance matrix for z_t has a maximum bandwidth of m for the first m rows and a bandwidth of q thereafter (Ansley, 1979), where the bandwidth is the number of nonzero elements in the row.

Let $\gamma_y(i)$ be the autocovariance of y_t at lag i, $\gamma_v(i)$ be the autocovariance of v_t at lag i, and $\gamma_{yv}(i)$ be the cross-covariance of y_t and v_t at lag i. Then,

$$\tilde{\mathbf{\Omega}}_{n,z} = \begin{bmatrix} \gamma_y(0) & \cdots & \cdots & \gamma_y(m) & & & & \\ \vdots & \ddots & & & \gamma_{vy}(q) & & 0 & \\ \vdots & & & \vdots & \vdots & \ddots & & \\ \gamma_y(m) & \cdots & \cdots & \gamma_y(0) & \gamma_{vy}(1) & \cdots & \gamma_{vy}(q) & \\ & \gamma_{vy}(q) & \cdots & \gamma_{vy}(1) & \gamma_v(0) & \ddots & \gamma_v(q) & \\ & & \ddots & \vdots & \vdots & \ddots & \ddots & \\ & & & \gamma_{vy}(q) & \vdots & \ddots & \ddots & \gamma_v(q) \\ & 0 & & & \gamma_v(q) & \ddots & \ddots & \vdots \\ & & & & & \gamma_v(q) & \cdots & \gamma_v(0) \end{bmatrix}_{n \times n} \tag{A.63}$$

Here we assume that $\theta^*(B) = \tilde{\theta}(B)$, since this is part of the hypothesis we are testing. However, to calculate $\mathbf{\Omega}^*_{n,z}$ we consider the covariance structure which would result from transforming a realization of the true process by a moving average polynomial which is not equal to the true moving average function. Define:

$$\tilde{v}_t = \frac{\hat{\phi}(B)}{\hat{\theta}(B)}\tilde{\theta}(B)a_t \tag{A.64}$$

Whereas the $\theta(B)$ polynomials in the numerator and denominator cancel in the case of v_t, we are no longer multiplying by the correct moving average function in (A.64), and therefore the cancellation does not occur. The expression for $\mathbf{\Omega}^*_{n,z}$ is:

$$\mathbf{\Omega}^*_{n,z} = \begin{bmatrix} \gamma^*_y(0) & \cdots & \gamma^*_y(m) & \gamma_{\tilde{v}_y}(m+1) & \cdots & \gamma_{\tilde{v}_y}(n) \\ \vdots & & \vdots & \vdots & \ddots & \\ \gamma^*_y(m) & \cdots & \gamma^*_y(0) & \gamma_{\tilde{v}_y}(1) & \cdots & \gamma_{\tilde{v}_y}(n-m) \\ \gamma_{\tilde{v}_y}(m+1) & \cdots & \gamma_{\tilde{v}_y}(1) & \gamma_{\tilde{v}}(0) & \cdots & \gamma_{\tilde{v}}(n-m+1) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{\tilde{v}_y}(n) & \cdots & \gamma_{\tilde{v}_y}(n-m) & \gamma_{\tilde{v}}(n-m+1) & \cdots & \gamma_{\tilde{v}}(0) \end{bmatrix}_{n \times n} \tag{A.65}$$

Note that while $\tilde{\mathbf{\Omega}}_{n,z}$ is banded, $\mathbf{\Omega}^*_{n,z}$ is not, in general. Also, $\tilde{\mathbf{\Omega}}^{-1}_{n,z}$ is not a banded matrix; however, Ma (1997) has proposed an expression for efficient computation of this inverse.

A.6 Profile Pair Sketches and Profile Traces

The information gathered over the course of profiling can be used to sketch nonlinear confidence regions for parameters or functions of parameters. To sketch a profile pair plot for two functions of parameters $g_1(\boldsymbol{\theta})$ and $g_2(\boldsymbol{\theta})$, follow the steps given in Figure A.5.

1. Profile $g_1(\boldsymbol{\theta})$.

2. Over the course of profiling $g_1(\boldsymbol{\theta})$, construct the matrix ${}^1\boldsymbol{M}$, where each row of ${}^1\boldsymbol{M}$ contains the results of the constrained optimization problem solved at each iteration of the profiling algorithm. Each column of the matrix contains values of one of the functions of parameters, one of the parameters, or τ, at each iteration. In table form, ${}^1\boldsymbol{M}$ is

Iteration	τ	$g_1(\bar{\boldsymbol{\theta}})$	---	$g_k(\bar{\boldsymbol{\theta}})$	$\bar{\theta}_1$	---	$\bar{\theta}_p$
1	${}^1m_{1,1}$	${}^1m_{1,2}$	---	${}^1m_{1,k+2}$	${}^1m_{1,k+3}$	---	${}^1m_{1,k+p+2}$
⋮							
${}^1h$	${}^1m_{{}^1h,1}$	${}^1m_{{}^1h,2}$	---	${}^1m_{{}^1h,k+2}$	${}^1m_{{}^1h,k+3}$	---	${}^1m_{{}^1h,k+p+2}$

   where $g_k(\boldsymbol{\theta})$ is the k^{th} function of parameters of interest, $\bar{\theta}_i$ is the i^{th} parameter of the vector of parameter values $\bar{\boldsymbol{\theta}}$, ${}^1h$ is the number of iterations needed to profile $g_1(\boldsymbol{\theta})$, and ${}^1m_{i,j}$ is the ij^{th} element of ${}^1\boldsymbol{M}$

3. Profile $g_2(\boldsymbol{\theta})$.

4. Over the course of profiling $g_2(\boldsymbol{\theta})$, construct the matrix ${}^2\boldsymbol{M}$ as for ${}^1\boldsymbol{M}$ in Step 2.

5. Using the data in ${}^1\boldsymbol{M}$, fit a spline curve $g_{\theta\tau,1}$ to τ as a function of $g_1(\boldsymbol{\theta})$. Also fit a spline $g_{\tau\theta,1}$ to $g_1(\boldsymbol{\theta})$ as a function of τ.

6. Using the data in ${}^2\boldsymbol{M}$, fit a spline curve $g_{\theta\tau,2}$ to τ as a function of $g_2(\boldsymbol{\theta})$. Also fit a spline $g_{\tau\theta,2}$ to $g_2(\boldsymbol{\theta})$ as a function of τ.

7. Use $g_{\theta\tau,1}$ to convert the $g_1(\bar{\boldsymbol{\theta}})$ column of ${}^2\boldsymbol{M}$ to a vector of τ values called τ_{12}.

8. Fit a spline $g_{\tau\tau,2}$ to τ_{12} as a function of the τ data from ${}^2\boldsymbol{M}$.

9. Use $g_{\theta\tau,2}$ to convert the $g_2(\bar{\boldsymbol{\theta}})$ column of ${}^1\boldsymbol{M}$ to a vector of τ values called τ_{21}.

10. Fit a spline $g_{\tau\tau,1}$ to τ_{21} as a function of the τ data from ${}^1\boldsymbol{M}$.

11. Use $g_{\tau\tau,1}$ to compute $q1$, the value of the spline at $\sqrt{rF(r,p-r;\alpha)}$, where r is the number of functions of parameters being considered jointly.

12. Use $g_{\tau\tau,1}$ to compute $q2$, the value of the spline at $-\sqrt{rF(r,p-r;\alpha)}$.

13. Use $g_{\tau\tau,2}$ to compute $p1$, the value of the spline at $\sqrt{rF(r,p-r;\alpha)}$.

14. Use $g_{\tau\tau,2}$ to compute $p2$, the value of the spline at $-\sqrt{rF(r,p-r;\alpha)}$. CONTINUED ON NEXT PAGE

Figure A.5: A step-by-step algorithm for sketching a profile pair plot for $g_1(\boldsymbol{\theta})$ and $g_2(\boldsymbol{\theta})$ (CONTINUED ON NEXT PAGE).

15. Let

$$\boldsymbol{sp} = \begin{bmatrix} 0 \\ \pi \\ acos\left(\frac{p1}{\sqrt{rF(r,p-r;\alpha)}}\right) \\ acos\left(\frac{p2}{\sqrt{rF(r,p-r;\alpha)}}\right) \end{bmatrix}$$

16. Let

$$\boldsymbol{sq} = \begin{bmatrix} acos\left(\frac{q1}{\sqrt{rF(r,p-r;\alpha)}}\right) \\ acos\left(\frac{q2}{\sqrt{rF(r,p-r;\alpha)}}\right) \\ 0 \\ \pi \end{bmatrix}$$

17. Let $\boldsymbol{a} = \frac{\boldsymbol{sp}+\boldsymbol{sq}}{2}$.

18. Let $\boldsymbol{d} = \boldsymbol{sp} + \boldsymbol{sq}$.

19. If any element of $\boldsymbol{d}$ is negative, change the sign of that element and the sign of the corresponding element of $\boldsymbol{a}$.

20. Let $\boldsymbol{a}^T = [\boldsymbol{a}^T - 2\pi, \boldsymbol{a}^T, \boldsymbol{a}^T + 2\pi]$.

21. Let $\boldsymbol{d}^T = [\boldsymbol{d}^T, \boldsymbol{d}^T, \boldsymbol{d}^T]$.

22. Let $\boldsymbol{S1} = \boldsymbol{a} + \boldsymbol{d}/2$.

23. Let $\boldsymbol{S2} = \boldsymbol{a} - \boldsymbol{d}/2$.

24. Fit a spline $g_{S2,S1}$ to $\boldsymbol{S2}$ as a function of $\boldsymbol{S1}$.

25. Choose a series of 100 equally spaced values from 0 to 2π. Let this vector of values be $\boldsymbol{sps}$.

26. Use $g_{S2,S1}$ to compute the vector $\boldsymbol{sqs}$ which corresponds to the values in $\boldsymbol{sps}$.

27. Let $\boldsymbol{\tau}_{\bullet 1} = cos(\boldsymbol{sps})\sqrt{rF(r,p-r,\alpha)}$.

28. Let $\boldsymbol{\tau}_{\bullet 2} = cos(\boldsymbol{sqs})\sqrt{rF(r,p-r,\alpha)}$.

29. Use $g_{\tau\theta,1}$ to convert $\boldsymbol{\tau}_{\bullet,1}$ to the vector $\boldsymbol{\theta}_{\bullet,1}$.

30. Use $g_{\tau\theta,2}$ to convert $\boldsymbol{\tau}_{\bullet,2}$ to the vector $\boldsymbol{\theta}_{\bullet,2}$.

31. Plot $\boldsymbol{\tau}_{\bullet,2}$ versus $\boldsymbol{\tau}_{\bullet,1}$.

Figure A.5: A step-by-step algorithm for sketching a profile pair plot for $g_1(\boldsymbol{\theta})$ and $g_2(\boldsymbol{\theta})$ (adapted from Bates and Watts (1988)).

A.7 Nomenclature

a_t	= white noise sequence
$\boldsymbol{a}$	= $p \times 1$ vector of constants
$A(q^{-1})$	= polynomial in the backshift operator q^{-1}
$B(q^{-1})$	= polynomial in the backshift operator q^{-1}
c	= a constant
$cov(\hat{\boldsymbol{\theta}})$	= variance covariance matrix of $\hat{\boldsymbol{\theta}}$
$C(q^{-1})$	= polynomial in the backshift operator q^{-1}
$\boldsymbol{e}$	= $n \times 1$ column vector of estimated random errors
$F(p, n-p; \alpha)$	= upper α quantile for the F distribution with p and $n-p$ degrees of freedom
$g(\boldsymbol{\theta})$	a function of parameters
$\hat{\dot{g}}$	= vector of derivative of $g(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ evaluated at $\hat{\boldsymbol{\theta}}$
$\boldsymbol{H}$	= Hessian matrix of $S(\boldsymbol{\theta})$
$L(\boldsymbol{\theta})$	= likelihood function evaluated at $\boldsymbol{\theta}$
$\mathcal{L}(\boldsymbol{\theta})$	= natural logarithm of the likelihood function of $\boldsymbol{\theta}$
LR	= likelihood ratio
$\mathcal{LR}$	= natural logarithm of the likelihood ratio
$LI(g(\boldsymbol{\theta}))$	= likelihood interval for $g(\boldsymbol{\theta})$
n	= number of observations
na, nb, nc	= orders of the polynomials $A(q^{-1})$, $B(q^{-1})$, $C(q^{-1})$, respectively
p	= number of estimated parameters
s	= estimated standard deviation of the random errors
$S(\boldsymbol{\theta})$	= sum of squared errors
$S^*(\boldsymbol{\theta})$	= sum of squared errors
se	= standard error
$t(n-p; \alpha/2)$	= upper $\alpha/2$ quantile of the t distribution with $n-p$ degrees of freedom
$v(u)$	= covariance at lag u
$\boldsymbol{V}$	= $n \times p$ matrix of elements v_{ij} representing the first
w_t	= a time series
x_t	= a time series
$\boldsymbol{x}$	= $1 \times m$ row vector of m independent variables
$\boldsymbol{X}$	= $n \times p$ matrix of elements x_{ij} representing the level of the j^{th} independent variable for observation i
y	= response variable
$\boldsymbol{y_n}$	= $n \times 1$ column vector of values of the response variable
z	= forward shift operator
z_t	= a time series transformed using Ansley's transformation derivative of $f(\boldsymbol{x_i}, \boldsymbol{\theta})$ with respect to the j^{th} parameter

Greek letters

α	= significance level
ϵ	= additive random error
$\boldsymbol{\epsilon}$	= $n \times 1$ column vector of random errors
θ_i	= i^{th} parameter of a model
$\boldsymbol{\theta}$	= $p \times 1$ vector of parameters
$\hat{\boldsymbol{\theta}}$	= $p \times 1$ vector of maximum likelihood estimates of the parameters
$\tilde{\boldsymbol{\theta}}$	= location of a constrained maximum of $L(\boldsymbol{\theta})$
λ	= a constant which defines the confidence level
$\nu(u)$	= covariance at lag u
$psi(t, \boldsymbol{\theta})$	= $p \times 1$ vector of derivatives of a_t with respect to $\boldsymbol{\theta}$
$\rho(u)$	= covariance at lag u
σ	= standard deviation
$\tau(g(\boldsymbol{\theta}))$	= profile t statistic for $g(\boldsymbol{\theta})$
Υ	= expected value of $\mathcal{I}$
ϕ_{kk}	= k^{th} partial autocorrelation
$\phi_{i,max}$	= upper stability/invertibility bound for the i^{th} parameter
$\chi^2(1)$	= the chi-squared distribution with 1 degree of freedom
$\boldsymbol{\Omega}_{n,z}$	= variance covariance matrix for z_t

Superscripts

$*$	= a true value
$\hat{}$	= a maximum likelihood estimate
$\tilde{}$	= a constrained estimate

Abbreviations

AR	autoregressive
ARMA	autoregressive moving average
ARMAX	autoregressive moving average with exogenous inputs
iid	independently and identically distributed
MLE	maximum likelihood estimate
PACF	partial autocorrelation function
SISO	single input single output