COUPLED AND SEPARABLE ITERATIONS
IN NONLINEAR ESTIMATION

G.K. Smyth

A thesis submitted to the
Australian National University
for the degree of Doctor of Philosophy

September, 1985
DECLARATION

Much of the material in chapters 5 to 8 is contained in Osborne and Smyth (1985) and (1986), and represents joint work with my supervisor Dr M.R. Osborne. With this qualification, and unless otherwise stated, the work in this thesis is my own.

G.K. Smyth


ACKNOWLEDGEMENTS

It is a pleasure to be able to thank some of those who have helped me on the long and sometimes winding road to completing this thesis. Firstly, the sustained enthusiasm, motivation and concrete help provided by Mike Osborne was critical. Robin Milne, Pat Moran, Sue Wilson, Ted Hannan and Chip Heathcote read early drafts of material which became part I. Pat Moran suggested the general topic of varying parameter models. Rob Womersley, Andreas Griewank, John Henstridge, Laimonis Kavalieris and Norm Campbell provided valuable comments and discussion.

Thanks also to Terry Speed and Robin Milne for moral support. To Norma Chin for very superior typing. To the Australian Government for financial support. And to Margaret for friendship and help at the eleventh hour.
GLOSSARY OF NOTATION

\( \dot{x} \) \hspace{1cm} \text{derivative}

\( x = (x_i) \) \hspace{1cm} \text{vector and components}

\( \|x\| \) \hspace{1cm} \text{Euclidean norm}

\( \|x\|_\infty \) \hspace{1cm} \text{supremum norm}

\( A = (a_{ij}) \) \hspace{1cm} \text{matrix and elements}

\( \|A\| \) \hspace{1cm} \text{matrix norm subordinate to Euclidean vector norm}

\( \|A\|_F \) \hspace{1cm} \text{Frobenius norm}

\( \text{tr } A \) \hspace{1cm} \text{trace}

\( \rho(A) \) \hspace{1cm} \text{spectral radius}

\( |A| \) \hspace{1cm} \text{determinant}

\( A^{\frac{1}{2}} \) \hspace{1cm} \text{Choleski factor}

\( \{d\} \) \hspace{1cm} \text{diagonal matrix}

\( 1 \) \hspace{1cm} \text{vector of ones}

\( \mathbf{1}_X \) \hspace{1cm} \text{coordinate vectors}

\( x^k \) \hspace{1cm} \text{sequence of vectors}

\( x^0 \) \hspace{1cm} \text{starting value}

\( \hat{x}^0 \) \hspace{1cm} \text{true value}

\( \hat{x} \) \hspace{1cm} \text{estimate}

\( \mathbb{R}(X) \) \hspace{1cm} \text{space spanned by columns of } X

\( \mathbb{E} \) \hspace{1cm} \text{expectation}

\( \mathbb{D} \) \hspace{1cm} \text{dispersion}

\( N(\mu, \sigma^2) \) \hspace{1cm} \text{normal distribution}

\( G(\alpha) \) \hspace{1cm} \text{gamma distribution}

\( \chi^2_P \) \hspace{1cm} \text{chi-square distribution}

\( \sim \) \hspace{1cm} \text{distributed as}

\( \approx \) \hspace{1cm} \text{asymptotically distributed as}

\( \approx \) \hspace{1cm} \text{asymptotically equal}

\( a.s. \) \hspace{1cm} \text{converges almost surely to}

\( \square \) \hspace{1cm} \text{end of proof}
ABSTRACT

This thesis deals with algorithms to fit certain statistical models. We are concerned with the interplay between the numerical properties of the algorithm and the statistical properties of the model fitted.

Chapter 1 outlines some results, concerning the construction of tests and the convergence of algorithms, based on quadratic approximations to the likelihood surface. These include the relationship between statistical curvature and the convergence of the scoring algorithm, separable regression, and a Gauss-Seidel process which we called coupled iterations.

Chapters 2, 3 and 4 are concerned with varying parameter models. Chapter 2 proposes an extension of generalized linear models by including a linear predictor for (a function of) the dispersion parameter also. Chapter 3 deals with various ways to go outside this extended generalized linear model framework for normally distributed data. Chapter 4 briefly describes how coupled iterations may be applied to autoregressive and multinormal models.

Chapters 5 to 8 apply a generalization of Prony's classical parametrization to solve separable regression problems which satisfy a linear homogeneous difference equation. Chapter 5 introduces the problem, specifies the assumptions under which asymptotic results are proved, and shows that the reduced normal equations may be expressed as a nonlinear eigenproblem in terms of the Prony parameters. Chapter 6 describes the algorithm which results from solving the eigenproblem, including some computational details. Chapter 7 proves that the algorithm is asymptotically stable. Chapter 8 compares the convergence of the algorithm with that of Gauss-Newton by way of simulations.
TABLE OF CONTENTS

Declaration ii
Acknowledgements iii
Glossary of Notation iv
Abstract v

Preamble 1

Chapter 1: Introduction
1.1 Outline 3
1.2 A word on notation 3
1.3 The likelihood function 5
1.4 Large sample tests 5
1.5 The scoring algorithm 6
1.6 Least squares and Gauss-Newton 7
1.7 Gauss-Newton and generalized linear models 8
1.8 Convergence 10
1.9 Curvature 12
1.9.1 Normal curvature 12
1.9.2 Bates and Watts intrinsic curvature 14
1.9.3 Efron curvature 19
1.10 Tests of composite hypotheses 20
1.11 Nested and coupled iterations 23
1.12 Convergence rates of partitioned algorithms 26
1.12.1 Nested iterations 26
1.12.2 Coupled iterations 27
1.13 Separable regression 28
1.14 Tests in submodels 30
1.15 Asymptotic stability of the Gauss-Newton iteration 31

PART I: COUPLED ITERATIONS AND VARYING PARAMETER MODELS

Preamble 36

Chapter 2: Generalized Linear Models with Varying Dispersion Parameters
2.1 Introduction 38
PART II: SEPARABLE REGRESSION AND PRONY'S PARAMETRIZATION

Preamble

Chapter 5: Prony's Parametrization

5.1 Introduction
5.2 The asymptotic sequence
5.3 Circulant matrices
5.4 Difference equations
5.5 The normal equations as a nonlinear eigenproblem
5.6 Recurrence equations
5.7 Differential equations
5.8 Exponential fitting
5.9 Rational fitting

Chapter 6: A Modified Prony Algorithm

6.1 Introduction
6.2 A modified Prony algorithm
   6.2.1 A sequence of linear eigenproblems
   6.2.2 Implementation
   6.2.3 Choice of scale
6.3 Calculation of B
   6.3.1 The general case
   6.3.2 Exponential fitting
   6.3.3 Rational fitting
6.4 Eigenstructure of B
6.5 Prony's method
6.6 More on exponential fitting
   6.6.1 Consistency
   6.6.2 Recovery of the rate constants
   6.6.3 Eigenstructure of $B_C$
6.7 Modified Prony with linear constraints
   6.7.1 Augmenting the Prony matrix
   6.7.2 Deflating the Prony matrix
   6.7.3 Exponential fitting

Chapter 7: Asymptotic Stability

7.1 Introduction
7.2 The convergence matrix
7.3 Invariance of the convergence factor
PREAMBLE

This thesis deals with algorithms to fit certain statistical models. We are concerned with the interplay between the numerical properties of the algorithm and the statistical properties of the model fitted. The author sees this work as being in the spirit of the papers of Nelder and Wedderburn (1972) and Jennrich (1969).

In their influential paper, Nelder and Wedderburn introduced the notion of a generalized linear model. They showed how the method of scoring, and its interpretation as a series of linear regressions, provides a unified treatment of the likelihood calculations. Generalized linear models suppose a distribution which is an exponential family plus a scale or dispersion parameter, and a linear predictor for a function of the mean of that distribution. In chapters 2 and 3 we propose an extension of generalized linear models by including a linear predictor for (a function of) the dispersion parameter also. We show that the resulting model may be fitted and analysed by fitting two simpler models in turn. The case of the normal distribution is examined in detail.

Similar ideas extend to certain dependent and multivariate models which are dealt with in chapter 4. Chapters 2, 3 and 4 comprise part I of this thesis.

In his 1969 paper, Jennrich proved the consistency and asymptotic normality of nonlinear least squares estimates and, under the same conditions, the asymptotic stability of the Gauss-Newton iteration.
We restrict our attention to separable regressions which satisfy exactly a homogeneous difference equation. For these regressions we can reparameterize in terms of the coefficients of the difference equation, and solve the normal equations using the algorithm of Osborne (1975). Under similar conditions to those of Jennrich (1969) we prove the asymptotic stability of the algorithm. The major emphasis is on the case of exponential fitting. Chapters 5 to 8 comprise part II.

In chapter 1 we outline some results, concerning the construction of tests and the convergence of algorithms, based on quadratic approximations to the likelihood surface. After this introductory chapter, parts I and II may be read independently.
CHAPTER 1
INTRODUCTION

1.1 Outline

In this chapter we outline some basic likelihood theory. Quadratic approximation to the likelihood surface leads to large sample statistical tests and to the scoring algorithm. In the case of fitting a mean to normal data, or solving a least squares problem, we specialize to the Gauss-Newton algorithm. Generalized linear models are developed as a family of likelihoods for which the scoring iteration takes the Gauss-Newton form. Sufficient conditions for local convergence of the scoring and Gauss-Newton algorithms are related to Efron's (1975) and Bates and Watts' (1980) measures of curvature. Partitioning the parameter vector leads on the one hand to tests of composite hypotheses, and on the other to modification of the scoring algorithm and to separable regression and coupled iterations in particular. Finally we outline Jennrich's (1969) proof of the asymptotic stability of the Gauss-Newton algorithm.

1.2 A word on notation

Mention needs to be made of several idiosyncratic notations. Firstly vector differentiation and multiplication of tensors. If $\lambda$ is a scalar function of $\theta \in \mathbb{R}^p$, then differentiation produces tensors

$$\dot{\lambda} = \frac{\partial \lambda}{\partial \theta}$$

and
\[ \dot{\mathbf{x}} = \left( \frac{\partial^2 x}{\partial \theta_i \partial \theta_j} \right) \]

which we take to be column vector and matrix respectively. If 
\( y(\theta) \in \mathbb{R}^n \) is a vector function of \( \theta \), then
\[ \dot{y} = \left( \frac{\partial y_i}{\partial \theta_j} \right) \]

and
\[ \ddot{y} = \left( \frac{\partial^2 y_i}{\partial \theta_j \partial \theta_k} \right) \]

are two and three dimensional tensors respectively. We follow the 
common (but not entirely consistent: differentiation now produces a 
row vector!) convention of identifying \( \dot{\mathbf{y}} \) with a \( n \times p \) matrix. No 
such representation can be made for \( \ddot{\mathbf{y}} \), but we will include it also 
in matrix expressions whenever it is clear which faces act on which. 
Thus, if \( \mathbf{y} \in \mathbb{R}^n \) and \( \mathbf{v} \in \mathbb{R}^p \), we will write
\[ \mathbf{u}^T \ddot{\mathbf{y}} = \left( \frac{n}{2} \frac{\partial y_i}{\partial \theta_j \partial \theta_k} \mathbf{y}_{j,k=1} \right)^p \]

for a \( p \times p \) matrix,
\[ \mathbf{v}^T \ddot{\mathbf{y}} = \left( \frac{p}{2} \frac{\partial^2 y_i}{\partial \theta_j \partial \theta_k} \mathbf{v}_{j,k=1} \right)^n \]

for an \( n \)-vector, and even
\[ \mathbf{z}^T \ddot{\mathbf{y}} = \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{\partial^2 y_i}{\partial \theta_j \partial \theta_k} \mathbf{v}_{j,k=1} \]

which is a scalar.

Secondly, scalar functions are taken to operate pointwise on 
vectors. For example, if \( \mathbf{x} \in \mathbb{R}^n \) then \( \mathbf{e}^T = (e^T) \) is also an 
\( n \)-vector. Similarly, if \( \mathbf{y} \in \mathbb{R}^n \) also, then \( \mathbf{x} \cdot \mathbf{y} = (x_i y_i) \).
represents the vector of componentwise products.

If \( x \in \mathbb{R}^n \) then \( \langle x \rangle \) represents the diagonal matrix with the \( x \) as diagonal elements. The symbols \( \mathbb{E} \) and \( \mathbb{D} \) are used for expectation and variance or dispersion respectively. Finally if \( X \) is a matrix then \( \mathcal{R}(X) \) is the linear space spanned by its columns.

1.3 The Likelihood function

Let \( \ell(\theta) \) be a log-likelihood function of vector parameter \( \theta \in \Theta \subseteq \mathbb{R}^p \) and (suppressed in our notation) data vector \( y \in \mathbb{R}^n \).

Write \( \nu(\theta) \) for the efficient score vector, \( \mathcal{I}(\theta) \) for the observed information, and \( I(\theta) = \mathbb{E}_{\theta}(\nu(\theta)) = \mathbb{D}_{\theta}(\nu(\theta)) \) for the observed information matrix, the expectation being taken at the same value of \( \theta \) as used to evaluate \( \nu \) and \( \mathcal{I} \).

We will generally assume conditions under which the maximum likelihood estimate \( \hat{\theta} \) is consistent, a form of the Central Limit Theorem can be applied to \( n^{-\frac{1}{2}} \mathcal{I} \), and a form of the Law of Large Numbers can be applied to \( \frac{1}{n} \mathcal{I} \). We will assume that \( I \) is nonsingular and is a good approximation to \( n^{-\frac{1}{2}} \mathcal{I} \) for \( \theta \) near the true value \( \theta_0 \).

Throughout this thesis we use methods based on \( I \) rather than \( -\mathcal{I} \) because it is the exact covariance matrix of \( \mathcal{I} \) and hence non-negative definite, and because it is algebraically simpler in the cases we consider. But see Efron and Hinkley (1978).

1.4 Large sample tests

The approximate normality of \( \mathcal{I} \) gives us immediately that

\[
(1.1) \quad I(\theta_0)^{-\frac{1}{2}} \mathcal{I}(\theta_0) \overset{\text{a}}{\sim} N(0, I)
\]
where \((I^{-\frac{1}{2}})^T I^{-\frac{1}{2}} = I^{-1}\). This is the score test statistic of the hypothesis \(\theta = \theta_0\), proposed by Rao (1947).

Expanding \(\hat{h}\) in a linear Taylor series about \(\theta_0\), and approximating \(-\hat{h}(\theta_0)\) by \(I(\theta_0)\), gives

\[
(1.2) \quad \hat{\theta} - \theta_0 \approx I(\theta_0)^{-1} \hat{h}(\theta_0)
\]

and hence

\[
(1.3) \quad I(\theta_0)^{-\frac{1}{2}}T (\hat{\theta} - \theta_0) \approx N(0, I).
\]

This is the Wald test statistic, proposed by Wald (1943).

Expanding \(\ell\) in a quadratic Taylor series about \(\hat{\theta}\), and approximating \(-\hat{h}(\hat{\theta})\) by \(I(\hat{\theta}_0)\), gives

\[
(1.4) \quad 2(\ell(\hat{\theta}) - \ell(\theta_0)) \approx (\hat{\theta} - \theta_0)^T I(\theta_0) (\hat{\theta} - \theta_0)
\]

and hence

\[
2(\ell(\hat{\theta}) - \ell(\theta_0)) \approx \chi^2.
\]

This is the likelihood ratio test statistic of Neyman and Pearson (1928).

Under standard conditions, all three large sample tests are asymptotically equivalent.

1.5 The scoring algorithm

Let \(\hat{\theta}^k\) be an approximation to that value of \(\hat{\theta}\) which maximizes \(\ell(\hat{\theta})\). We may seek to improve that approximation by maximizing the quadratic expansion of \(\ell\) about \(\hat{\theta}^k\). This is the classical Newton-Raphson algorithm. Approximating \(-\hat{h}(\hat{\theta}^k)\) by \(I(\hat{\theta}^k)\) results in the update
which defines an iteration of the scoring algorithm, usually attributed to Fisher (1925). If the process converges then repeated iteration yields a stationary point which satisfies \( \hat{\lambda}(\hat{\theta}) = 0 \), which we take to be the maximum likelihood estimate.

Note the similarity between (1.2) and (1.5). In fact the score test is equal to the Wald test with \( \hat{\theta} \) replaced by the value obtained after one scoring iteration from \( \hat{\theta}_0 \). One scoring iteration from a consistent estimator produces an estimator which may be asymptotically efficient even if the maximum likelihood estimate is not, and is equivalent to it if it is. (Le Cam (1956); Bickel (1975) proved similar results for one-step Huber estimators.)

### 1.6 Least squares and Gauss-Newton

Suppose the data is normal with mean vector \( \mu \), a function of \( \beta \in \mathbb{R}^p \), and covariance matrix \( \Sigma \) known up to the multiplier \( \sigma^2 \). Then

\[
(1.6) \quad \ell = -\frac{1}{2} \log |\Sigma| - \frac{1}{2} \sigma^{-2} (\bar{y} - \mu(\beta))^{T} \Sigma^{-1} (\bar{y} - \mu(\beta))
\]

so

\[
\hat{\beta}_\ell = \sigma^{-2} \mu^{T} \Sigma^{-1} (\bar{y} - \mu) \]

\[
\hat{\beta}_I = -\sigma^{-2} \mu^{T} \Sigma^{-1} \mu + \sigma^{-2} \mu^{T} \Sigma^{-1} (\bar{y} - \mu)
\]

and

\[
I_\beta = \sigma^{-2} \mu^{T} \Sigma^{-1} \mu.
\]

The value of \( \sigma^2 \) does not affect the scoring iteration for \( \hat{\beta} \), which is

\[
(1.7) \quad \hat{\beta}^{k+1} - \hat{\beta}^k = (\mu^{T} \Sigma^{-1} \mu)^{-1} \mu^{T} \Sigma^{-1} (\bar{y} - \mu)
\]
In this case the scoring iteration is known as the **Gauss-Newton algorithm**. In the simplest case $y$ is linear in $\beta$ and we have weighted linear regression. The quadratic approximation to $\ell$ is then exact, $I + \lambda = 0$, and Gauss-Newton converges in one step from any starting point.

Maximizing the log-likelihood (1.6) with respect to $\beta$ is equivalent to minimizing the sum of squares

$$\phi(\beta) = (y - \mu(\beta))^T V^{-1}(y - \mu(\beta)).$$

In fact the algorithm (1.7) depends on the distribution of $y$ only through its first and second moments. It is generally applied to least squares problems when only second moment assumptions are made. Indeed it is formally available whenever the objective function can be written as a sum of squares. To minimize

$$f(\beta)^T f(\beta)$$

the algorithm takes the form

$$\beta^{k+1} - \beta^k = (fTf)^{-1} fTf.$$

This process may perform poorly if $\mathbb{E}(f) \neq 0$; the application of Ross (1982) suffers from this defect.

1.7 Gauss-Newton and generalized linear models

If we allow that the weight matrix $V$ may depend on $\beta$, the most general objective functions $k(\beta)$ for which the scoring iteration takes the form (1.7) are those for which

$$k = \sigma^{-2} \mu(\beta)^T V(\beta)^{-1}(y - \mu(\beta)).$$
These are quasi-likelihoods of Wedderburn (1974) and McCullagh (1983).

If $k$ is a log-likelihood it must have the form

$$\sigma^{-2}\{y^T\psi(\beta) - b(\psi(\beta))\} + c(y, \sigma)$$

for suitably chosen functions $b$ and $c$, and for $\psi$ a transformation of $\mu$. In fact it must be that

$$\mathbb{E}(\psi) = \psi(\beta) = b(\psi)$$

and

$$\mathbb{D}(\psi) = \sigma^2 \psi(\beta) = \sigma^2 b(\psi).$$

Now let us restrict the $y_i$ to be independent, and to have densities of the same form except for possibly different values of $V_i$ and known weights $w_i$. Then the log-likelihood (1.9) becomes

$$\sigma^{-2} \sum_i \{w_i(y_i V_i - b(V_i)) + c(y_i, w_i \sigma^{-2})\}$$

where $b$ and $c$ now take scalar arguments. Let us further assume that the dependence of $\mu$ on $\beta$ takes the form

$$g(\mu) = X\beta$$

where $X$ is an $n \times p$ matrix and $g$ a scalar function operating pointwise on $\mu$. Then (1.10) and (1.11) define Nelder and Wedderburn's (1972) generalized linear models. The vector $X\beta$ is known as the linear predictor, $g$ as the link function and $\sigma^2$ as the dispersion parameter.

The matrix $V$ is now diagonal, and equal to $(v(\mu)^{-1} \omega)$ with $v(\mu) = b'(\psi)$. And from (1.11) we see that $\mu = (g(\psi)^{-1} X$. For generalized linear models the scoring iteration is therefore

$$\beta^{k+1} - \beta^k = (X^T (g^{-2} \psi^{-1} \omega) X)^{-1} X^T (g^{-2} \psi^{-1} \omega) (y - \mu)$$
with the right hand side evaluated at $\beta^k$. A little rearrangement puts (1.12) in the form of a linear regression estimate

\begin{equation}
\beta^{k+1} = (X'WX)^{-1}X'Wz
\end{equation}

with

$$W = \langle g^2 \rangle^{-1} w$$

and

$$z = \langle g \rangle (y - \mu) + g(\mu).$$

Iterations of the form (1.13) are often called \textit{iteratively reweighted least squares}. The relationship between reweighted least squares, Gauss-Newton and maximum likelihood estimation is further discussed by Jorgenson (1983) and Green (1984). McCullagh and Nelder (1983) is a recent monograph on generalized linear models; see also the review by Pregibon (1984). Fahrmeir and Kaufmann (1985) prove the consistency and asymptotic normality of maximum likelihood estimates for generalized linear models. Existence and uniqueness results are collected by Wedderburn (1976).

In practice Gauss-Newton is applied with line searches or trust regions to secure convergence; commonly this is accomplished by the Levenberg-Marquardt modification (Fletcher, 1980). A similar modification for generalized linear models has been described by Osborne (1985).

1.8 Convergence

Consider the iterative process defined by

$$x^{k+1} = F(x^k).$$

A point $x^*$ is said to be a \textit{point of attraction} if the process
converges to $\mathbf{x}^*$ from any starting point in some neighbourhood of $\mathbf{x}^*$. In that case the process is said to be stable at $\mathbf{x}^*$. A sufficient condition for stability is given by Ostrowski's Theorem (Ortega and Rheinboldt, 1970), namely that $\mathbf{F}$ is (Frechet) differentiable at $\mathbf{x}^*$ and that

$$R = \rho(\mathbf{F}(\mathbf{x}^*)) < 1$$

where $\rho(*)$ is spectral radius. We call $\mathbf{F}(\mathbf{x}^*)$ the convergence matrix, and $R$ the convergence factor, since if $R > 0$ then convergence is linear and ultimately $\|\mathbf{x}^{k+1} - \mathbf{x}^*\|/\|\mathbf{x}^k - \mathbf{x}^*\| = R$. If $R$ is achieved by a positive eigenvalue of $\mathbf{F}$ then convergence is ultimately monotonic, if by a negative eigenvalue then convergence is ultimately oscillatory.

Differentiation of (1.5) at $\hat{\mathbf{z}}$ gives the convergence matrix of the scoring algorithm

$$G = I^{-1}(\hat{\mathbf{z}} + I)$$

This specializes to

$$G = (\hat{\mathbf{u}}^T\hat{\mathbf{u}})^{-1} \hat{\mathbf{u}}^T(\hat{\mathbf{y}} - \hat{\mathbf{u}})$$

for Gauss-Newton with identity weight matrix.

Note that (1.14) has Rayleigh quotient

$$1 + \frac{\mathbf{z}^T \mathbf{z}}{\mathbf{z}^T I \mathbf{z}} \mathbf{z}^T \mathbf{z}$$

If $\mathbf{z}$ is negative definite, as it should be at a maximum of $\mathbf{z}$, then the quotient is less than one for all $\mathbf{z}$. Therefore if the scoring

---

$\dagger$ Actually $R$ is the root-convergence factor which is possibly different from the quotient-convergence factor. See Ortega and Rheinboldt (1970).
algorithm diverges from close to a maximum, it is likely to do so in an oscillatory manner.

Consider a reparameterization with Jacobian \( J \). Then \( I \) transforms to \( J^T I J \) and \( \dot{x}(\tilde{\theta}) \) to \( J(\tilde{\theta})^T \dot{x}(\tilde{\theta}) J(\tilde{\theta}) \). Hence (1.14) transforms to

\[
J^{-1} I^{-1} (\ddot{x} + I) J
\]

to which it is similar. The spectral radius of (1.14) then is a geometric invariant; it does not depend on the parameterization.

If the Gauss-Newton algorithm is actually implemented in a modified form its convergence matrix will be more complicated than (1.15). The behaviour of the unmodified algorithm though, will be relevant sufficiently close to a solution.

1.9 Curvature

1.9.1 Normal curvature

The eigenvalues of (1.14) and (1.15) may be given useful geometric interpretations. In this section and the next we restrict our attention to (1.15), and return to (1.14) in §1.9.3. In fact (1.15) is similar to the symmetric matrix

\[
B = (\mu^T \mu)^{-1/2} \mu^T (y - \mu) (\mu^T \mu)^{-1/2}
\]

which is the "effective residual curvature matrix" of Hamilton, Watts and Bates (1982). Before exploring this further we discuss the curvatures of one-dimensional curves. Unless otherwise indicated, \( \mu \) and its derivatives are always evaluated at \( \hat{\theta} \).

Let \( f \) be a function mapping \( \mathbb{R} \) into \( \mathbb{R}^n \). The range of \( f \) then defines a one-dimensional curve in \( n \)-space. Consider the
limiting circle through the points \( f(a - \varepsilon) \), \( f(a) \) and \( f(a + \varepsilon) \) as \( \varepsilon \to 0 \). The normal curvature at \( a \) may be defined to be the inverse radius of this limiting circle, and can be calculated as

\[
\kappa = \frac{\| T \cdot P_N f \|}{\| f \|}
\]

where \( P_N \) is the projection onto \( N = R(\dot{f}(a)) \) (Johansen, 1984, pp. 80-81).

Now consider one-dimensional curves on the solution locus \( \{ y(0) : 0 \in \Theta \} \). Corresponding to any direction \( v \in \mathbb{R}^p \), there is a line \( \Theta(\alpha) = \hat{\Theta} + \alpha v \) from \( \hat{\Theta} \) in the parameter space, and a lifted one-dimensional curve on the solution locus defined by \( \dot{y}(\Theta(\alpha)) \). We have that

\[
d\mu = v
\]

and

\[
d^2\mu = v T \cdot T \cdot v.
\]

Furthermore the projection of \( v T \cdot \mu \) onto \( R(\dot{\mu}) \) is the same as its projection onto \( R(\dot{\mu}) \). So the normal curvature of the lifted curve is

\[
(1.16) \quad \kappa(v) = \frac{\| v T \cdot P_N \mu \|}{v T \cdot T \cdot v}
\]

where \( P_N \) is now the projection onto \( R(\dot{\mu}) \). (In the terminology of differential geometry, \( P_N \dot{\mu} \) is the second fundamental form of the surface \( \mu(\Theta) \), the information matrix \( T \cdot \mu \) being the first (Reed, 1975, Johansen, 1984).)

An essentially equivalent derivation of (1.16) would have arisen had we defined the one-dimensional curve as the intersection of \( R(\mu) \)
with the solution locus, that is as a normal cut. That approach would exhibit $\kappa$ as a function of the tangent direction $\mu v$. As a function of $\mu v$, $\kappa$ is a geometric invariant.

Now return to (1.15). The eigenvalues of $G$ are the stationary values of the Rayleigh quotient

$$q(\nu) = \frac{\nu^T (\gamma - \mu) \gamma}{\nu^T \gamma}.$$  

Let $e = (\gamma - \mu)/\|\gamma - \mu\|$, and let $P_e$ be the projection onto $R(e)$. Then

$$q(\nu) = \frac{\|\nu^T P_e \nu\|}{\nu^T \mu \nu}.$$  

Let $\lambda_1 < ... < \lambda_p$ be the eigenvalues of $G$ with eigenvectors $x_1, ..., x_p$. We see that $\lambda_i$ is the normal curvature at $a = 0$ of the curve $\mu(\hat{\theta} + \alpha x_i)$ imbedded in the space spanned by $\mu$ and $e$. We call the $\lambda_i$ the normal curvatures of the solution locus at $\hat{\theta}$ relative to $e$.

Hamilton, Watts and Bates (1982) showed that the $\lambda_i$ may be used to give the relative lengths of the axes of ellipsoidal likelihood or confidence regions. In particular, for likelihood regions, the relative lengths are $\nu_1, ..., \nu_p$ with $\nu_i = (1 - \lambda_i)^{-\frac{1}{2}}$.

1.9.2 Bates and Watts intrinsic curvature

The quantity (1.16) is the intrinsic curvature, in the direction associated with $\nu$, of Bates and Watts (1980). For the purposes of calibration, Bates and Watts (1980) define the relative intrinsic curvature.
Y(v) = \partial_p \frac{1}{2} \kappa(v) .

We now discuss the relationship between intrinsic curvature and the normal curvatures.

Consider the variability of (1.17) over data sets giving rise to the same least squares estimate \( \hat{\theta} \). If we assume \( \hat{\theta} \) to be the true value, the conditional distribution of \( Y - \mu(\hat{\theta}) \) given \( \hat{\mu}(\hat{\theta})^T (Y - \mu(\hat{\theta})) = 0 \) is \( N(0, \sigma^2_p N) \). Under this distribution

\[
\mathbb{E}_{Y|\hat{\theta}}(q^2(Y)) = \sigma^2 \frac{\|T_p \tilde{\mu}\|^2}{(\tilde{\mu}^T \mu)^2} = p^{-1} \gamma(Y) .
\]

In any particular direction then, \( p^{-1} \gamma(Y) \) gives the standard deviation of the Rayleigh quotient \( q(Y) \), conditional on the least squares estimate being at the point of evaluation. It may be viewed as a before-the-data estimate of the size of the Gauss-Newton convergence matrix in that direction and at that point.

Bates and Watts (1980) define their summary measures of intrinsic curvature by maximizing or averaging \( \gamma^2(Y) \) over \( Y \). Thus

\[
\Gamma^2 = \max_Y \gamma^2(Y)
\]

and

\[
\gamma_{RMS}^2 = \mathbb{E}_Y \gamma^2(Y)
\]

the expectation being taken over \( Y \sim N(0, (\mu^T \mu)^{-1}) \). The second measure, \( \gamma_{RMS}^2 \) is four times the measure of nonlinearity, \( N_\phi \), derived by Beale (1960). It can be calculated explicitly as follows. Let \( A_t \) be the \( t \)'th \( p \times p \) face of \( (\mu^T \mu)^{-1} T_p \mu \mu^T (\mu^T \mu)^{-1} \), and let \( \tilde{d} = (\mu^T \mu)^{-1/2} Y \). Then (1.16) becomes

\[
\kappa^2(\tilde{d}) = \frac{\sum (d^T A_t d)^2}{(d^T \tilde{d})^2} .
\]
Taking expectation over \( d \sim N(0, I) \), and using the independence of \( d^T A_t d \) and \( d^T d \) to carry the expectation through to numerator and denominator,

\[
\frac{1}{p^2} \gamma_{\text{RMS}}^2 = \frac{\sum_t E_d (d^T A_t d)^2}{\sum_t E_d (d^2)^2} = \frac{\sum_t \left\{ 2 \sum_i \lambda_{t,i}^2 + (\sum_i \lambda_{t,i}^2) \right\}}{2p+p^2}
\]

where \( \lambda_{t,1}, \ldots, \lambda_{t,p} \) are the eigenvalues of \( A_t \),

\[
\sum_t \left\{ 2 \| A_t \|_F^2 + (\text{tr } A_t)^2 \right\}
\]

\[
\frac{t}{p(p+2)}
\]

where \( \| \cdot \|_F \) is Frobenius norm and \( \text{tr}(\cdot) \) is trace.

Both summary measures inherit geometric invariance from \( K(y) \):

\( \gamma^2 \) is obtained by maximizing \( p^2 K(y) \) over tangent directions \( \hat{y} \);

\( \gamma_{\text{RMS}}^2 \) is obtained by averaging over \( \hat{y} \) symmetrically distributed in the tangent plant \( R(\hat{y}) \).

Both \( \gamma^2 \) and \( \gamma_{\text{RMS}}^2 \) provide useful bounds for the expected size of the convergence factor. In fact, since

\[
\max_v \mathbb{E} (q^2(\hat{y})) \leq \mathbb{E} (\max_v q^2(\hat{y}))
\]

we have immediately that

\[
\frac{1}{p} \gamma_{\text{RMS}}^2 \leq \frac{1}{p} \gamma^2 \leq \mathbb{E} \left( \rho^2(G) \right).
\]

We view \( \gamma^2 \) itself as a conservative (large) estimate of

\[ \mathbb{E}_y \rho^2(G) \]. Comparing \( \gamma^2 \) with \( F(p; n-p; .95)^{-1} \) as do Bates and Watts (1980) and Hamilton, Watts and Bates (1982), is seen to be very
We can also show that

\[ \frac{1}{p} \gamma^2_{\text{RMS}} = \frac{2E_{\gamma|\hat{\theta}} (\|G\|_F^2 + E_{\gamma|\hat{\theta}} (\text{tr} G)^2}{p(p+2)} \leq \frac{1}{p} E_{\gamma|\hat{\theta}} (\|G\|_F^2). \]

The equality of (1.18) is established as follows. Since 
\[(\gamma - \mu) \sim N(0, \sigma^2 P_N),\ (\gamma - \mu) \text{ can be written as } P_N \varepsilon \text{ with } \varepsilon \sim N(0, \sigma^2 I) .\] Then \(G\) has the same eigenvalues as \(\Sigma \varepsilon_t A_t\).

Therefore

\[ E_{\gamma|\hat{\theta}} (\text{tr} G) = \sigma^2 \Sigma \text{tr} A_t \]

and

\[ E_{\gamma|\hat{\theta}} (\|G\|_F^2) = \sigma^2 \Sigma \|A_t\|_F^2. \]

The inequality of (1.18) follows from the fact that \(\|G\|_F^{-2} (\text{tr} G)^2\) achieves its maximum of \(p\) when all the eigenvalues of \(G\) are equal.

Alternatively we could bound the expected square of the convergence factor using \(S_2 = E_{\gamma|\hat{\theta}} (\|G\|_F^2)\) itself, by

\[ \frac{1}{p} S_2 \leq E_{\gamma|\hat{\theta}} (\rho^2(G)) \leq S_2. \]

Higher moments may be bounded similarly. For example

\[ S_4 = E_{\gamma|\hat{\theta}} \text{tr}(G^4) = 3\sigma^4 \left\{ \Sigma \text{tr}(A_t^4) + 2 \Sigma \text{tr}(A_s A_t^2) \right\} \]

and

\[ \frac{1}{p} S_4 \leq E_{\gamma|\hat{\theta}} (\rho^4(G)) \leq S_4. \]

We conclude this section with a table extracted from table 1 of Hamilton, Watts and Bates (1982).† The table shows that \(\Gamma^2\) is

† Table 1 of Hamilton, Watts and Bates (1982) and Table 2 of Bates and Watts (1980) are not mutually consistent. Values of \(\Gamma\) were
Table 1. Maximum relative intrinsic curvature, minimum and maximum eigenvalues of the convergence matrix, and the convergence factor for 18 data sets.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>$\Gamma$</th>
<th>$\lambda_{-1}(G)$</th>
<th>$\lambda_p(G)$</th>
<th>$\rho(G)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.03</td>
<td>-.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>2</td>
<td>.06</td>
<td>-.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>3</td>
<td>.08</td>
<td>-.17</td>
<td>.00</td>
<td>.17</td>
</tr>
<tr>
<td>4</td>
<td>.07</td>
<td>-.10</td>
<td>.00</td>
<td>.10</td>
</tr>
<tr>
<td>5</td>
<td>.21</td>
<td>-.00</td>
<td>.10</td>
<td>.10</td>
</tr>
<tr>
<td>9</td>
<td>.18</td>
<td>-.06</td>
<td>.02</td>
<td>.06</td>
</tr>
<tr>
<td>13</td>
<td>.01</td>
<td>-.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>14</td>
<td>.15</td>
<td>-.00</td>
<td>.08</td>
<td>.08</td>
</tr>
<tr>
<td>15</td>
<td>.04</td>
<td>-.04</td>
<td>.00</td>
<td>.04</td>
</tr>
<tr>
<td>16</td>
<td>.04</td>
<td>-.00</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>17</td>
<td>.25</td>
<td>-.00</td>
<td>.15</td>
<td>.15</td>
</tr>
<tr>
<td>18</td>
<td>.00</td>
<td>-.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>19</td>
<td>.02</td>
<td>-.02</td>
<td>.00</td>
<td>.02</td>
</tr>
<tr>
<td>20</td>
<td>.02</td>
<td>-.06</td>
<td>.00</td>
<td>.06</td>
</tr>
<tr>
<td>21</td>
<td>.90</td>
<td>-.06</td>
<td>.26</td>
<td>.26</td>
</tr>
<tr>
<td>22</td>
<td>.08</td>
<td>-.19</td>
<td>.00</td>
<td>.19</td>
</tr>
<tr>
<td>23</td>
<td>.09</td>
<td>-.02</td>
<td>.10</td>
<td>.10</td>
</tr>
<tr>
<td>24</td>
<td>.37</td>
<td>-.04</td>
<td>.10</td>
<td>.10</td>
</tr>
</tbody>
</table>

credible as a conservative estimate of $\mathbb{E}(p^2(G))$, and that $\rho(G)$ is well within the bounds for a point of attraction for all 18 data sets.

For each of Bates and Watts' measures of intrinsic curvature there is a corresponding measure of parameter effects curvature which depends on the parametrization (Bates and Watts 1980, 1981; Kass, 1984). We have neglected these because $\rho(G)$ is a geometric

obtained by multiplying $\Gamma F(p; \nu; \alpha)$ from the former table by $F(p; \nu; \alpha)^{-1}$ from the latter table.

1.9.3 Efron curvature

In our discussion of curvature so far we have restricted ourselves to least squares problems and to particular directions in the parameter space. In the general case we consider the variability of (1.14) conditional on \( \hat{\delta} \). The covariances of the \( p^2 \) elements of the matrix \( G \) form a four-dimensional tensor

\[
\text{Cov}(\tilde{G}_{ij}, \tilde{G}_{mn} | \hat{\delta}) = (\gamma_{ij,mn})
\]

which holds all the second moment information about \( G \). This is the multiparameter counterpart of Efron's (1975) statistical curvature.

Let

\[ \text{Cov}(\tilde{\delta}_{ij}, \tilde{\delta}_{mn} | \hat{\delta}) = \alpha_{ij,mn} . \]

Then

\[ \gamma_{kj,m\ell} = \sum_{in} I_{kn} I_{\ell n} \alpha_{ij,mn} \]

where \( (I^{ij}) = I^{-1} \). Dividing again by \( I \) gives

\[ B_{k\ell} = \sum_{jm} \gamma_{kj,m\ell} I_{jm} \]

which is the order \( n^{-2} \) correction to the covariance matrix of the maximum likelihood estimate due to statistical curvature. See Reeds (1975), Madsen (1979), Amari (1982). Bounds for the expected squared convergence rate may be had from

\[ (1.20) \quad \text{E} y^2 | \hat{\delta} (\| G \|^2_2) = \sum_{ij} \gamma_{ij,ii,jj} . \]
Specializing to least squares gives

\[ \gamma_{mi, kn} = \gamma_0^2 \sum_{st} \sum_{jk} (\mu^T \mu)^{mj} \bar{\mu}_{s, ij} P^N_{st} \bar{\mu}_{t, kj} (\mu^T \mu)^{kn} \]

and (1.20) becomes (1.19).

Efron's statistical curvature is also discussed by Efron (1978) and Efron and Hinkley (1978).

1.10 Tests of Composite Hypotheses

Suppose that \( \theta \) is partitioned into \( \theta_1 \in \mathbb{R}^{p_2} \) and \( \theta_2 \in \mathbb{R}^{p_2} \), \( p_1 + p_2 = p \), and that

\[ \hat{\lambda} = \begin{pmatrix} \hat{\lambda}_1 \\ \hat{\lambda}_2 \end{pmatrix} \]

and

\[ I = \begin{pmatrix} I_1 & I_{12} \\ I_{21} & I_2 \end{pmatrix} \]

are conformal partitions of \( \hat{\lambda} \) and \( I \). Frequent use will be made in this and the next two sections of the block \( LD L^T \) decomposition

\[ (1.21) \quad I = I_{21} I_1^{-1} I \begin{pmatrix} I_1 & 0 \\ 0 & I_2 - I_{21} I_1^{-1} I_{12} \end{pmatrix} \begin{pmatrix} I & I_1^{-1} I_{12} \\ 0 & I \end{pmatrix}. \]

This decomposition suggests the transformation to orthogonal parameters

\[ (1.22) \quad L^T \theta = \begin{pmatrix} \theta_1 + I_1^{-1} I_{12} \theta_2 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \theta_{1.2} \\ \theta_2 \end{pmatrix} \]

with score vector
(1.23) \[ L^{-1} \mathbf{k} = \begin{bmatrix} \mathbf{k}_1 \\ \mathbf{k}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{k} \\ \mathbf{k}_2.1 \end{bmatrix} \]

and information matrix

(1.24) \[ L^{-1} I L^{-T} = \begin{bmatrix} I_1 & 0 \\ 0 & I_2-I_2 I_1^{-1} I_2 \end{bmatrix} = \begin{bmatrix} I_1 & 0 \\ 0 & I_2.1 \end{bmatrix}. \]

Suppose that we wish to test \( \theta_2 = \theta_{20} \), and the \( \theta_1 \) enter as unknown nuisance parameters. Locally all the information about \( \theta_1 \) and \( \theta_2 \) is contained in \( \mathbf{k}_1 \) and \( \mathbf{k}_2 \). So we form the conditional distribution of \( \mathbf{k}_2 \) given \( \mathbf{k}_1 \)

\[ \mathbf{k}_2 | \mathbf{k}_1 \overset{\mathcal{D}}{\sim} N(I_2 I_1^{-1} \mathbf{k}_1, I_2.1) \]

and obtain the statistic

(1.25) \[ I_{2.1}^{-\frac{1}{2}} \mathbf{k}_{2.1} \overset{\mathcal{D}}{\sim} N(0, I) \]

based on \( \mathbf{k}_2 \) but corrected for regression on \( \mathbf{k}_1 \).

It remains to replace \( \theta_1 \) with an estimate in (1.25). It is usual to use \( \hat{\theta}_1(\theta_{20}) \), the value of \( \theta_1 \) which maximizes \( \mathbf{k} \) with \( \theta_2 \) fixed at \( \theta_{20} \). Then (1.25) becomes

(1.26) \[ I_{2.1}^{-\frac{1}{2}} \mathbf{k}_{2.1}(\hat{\theta}_1(\theta_{20}), \theta_{20}) \overset{\mathcal{D}}{\sim} N(0, I) \]

which is Rao's (1948) score test for composite hypotheses. Note that at \( \theta_1 = \hat{\theta}_1(\theta_{20}) \), \( \mathbf{k}_{2.1} = \mathbf{k}_2 \) since \( \mathbf{k}_1 = 0 \). A statistic equivalent to (1.26) was independently derived by Aitchison and Silvey (1958, 1960), and has become entrenched in the econometrics literature as the \textit{Lagrange multiplier test}. See the review of Breusch and Pagan (1980).

\[ \dagger \] This is intended to be informal. But in fact locally or asymptotically \( \mathbf{k} \) is a sufficient statistic for \( \theta \).
Note that estimates need to be obtained under the null hypothesis only.

If some other estimate of $\theta_1$ is used, then (1.25) is the $C(\alpha)$ test proposed by Neyman (1959) (extended to multi-parameter hypotheses by Bühler and Puri (1966); see also Moran (1970)). Neyman's contribution was to show that the asymptotic properties of (1.25) are retained for any estimator which is "root-n consistent", that is for which $n^{1/2}|\hat{\theta}_1 - \theta_{10}|$ remains bounded in probability.

Partitioning (1.2) gives

\[(1.27) \quad \hat{\theta}_1 - \theta_{10} \overset{\text{a}}{\sim} I^{-1}\lambda 2.1 \lambda 2.1\]

which with (1.25) gives

\[(1.28) \quad I^{1/2} I_2.1 (\hat{\theta}_1 - \theta_{10}) \overset{\text{a}}{\sim} N(0, I)\]

which is the Wald test of a composite hypothesis. It is usual to use $\hat{\theta}_1$ for $\theta_1$ in (1.28) so that only unrestricted estimates are required.

Note that if $I^{1/2}$ is chosen to be the block Choleski factor

\[(1.29) \quad I^{1/2} = \begin{bmatrix} I & 0 \\ I_2.1 I_1^{-1} & I \end{bmatrix} \begin{bmatrix} I^{1/2} & 0 \\ 0 & I^{1/2} \end{bmatrix}\]

then (1.26) and (1.28) are simply the conformal components of (1.1) and (1.3).

We can decompose (1.4) also as

\[2\{\ell(\hat{\theta}_1, (\hat{\theta}_2 - \theta_{2.0}, 0^{(\theta_{20}))}) - \ell(\theta_1, (\hat{\theta}_2 - \theta_{20}))\} + 2\{\ell(\hat{\theta}_1, (\hat{\theta}_1, (\hat{\theta}_2 - \theta_{20})) - \ell(\theta_1, (\hat{\theta}_1, (\hat{\theta}_2 - \theta_{20}))\} \]

\[(1.30) \quad \overset{\text{a}}{\sim} (\hat{\theta}_1 - \theta_{10.20})^T I_1 (\hat{\theta}_1 - \theta_{10.20}) + (\hat{\theta}_2 - \theta_{20})^T I_2.1 (\hat{\theta}_2 - \theta_{20}) .\]

Identifying the second terms on the left and right hand sides of
as the terms conditional on \( \hat{\theta}_1 \) establishes

\[
(1.31) \ 2 \{ \ell(\hat{\theta}_1, \hat{\theta}_2) - \ell(\hat{\theta}_1, \hat{\theta}_{20}) \} \overset{\Delta}{=} (\hat{\theta}_2 - \theta_{20})^T I_{2,1} (\hat{\theta}_2 - \theta_{20}) \sim \chi^2_{p_2}.
\]

This is the likelihood ratio test statistic for the composite hypothesis. Note that both restricted \( \hat{\theta}_1(\theta_{20}) \) and unrestricted \( \hat{\theta}_1, \hat{\theta}_2 \) estimates are required.

The above treatment in terms of a partition of \( \hat{\theta} \) implies the form of the test statistics for general composite hypotheses specified in terms of restrictions. Suppose we wish to test that \( h(\theta) = 0 \), where \( h \) is a \( r \)-dimensional function of \( \theta \). If \( h \) has full rank then we define a reparametrization of \( \theta \) to \( \phi \), with \( \phi^T = (\phi_1^T, \phi_2^T) \) and \( \phi_2 = h \), and test \( \phi_2 = 0 \).

1.11 Nested and Coupled Iterations

Suppose that \( \hat{\theta} \) is partitioned as in §1.10. Using the decomposition (1.21) of \( I \) we may write the scoring iteration (1.4) as

\[
\begin{align*}
\theta_{1}^{k+1} - \theta_{1}^{k} &= \mathcal{F}_1(\theta_{1}^{k}, \theta_{2}^{k}) = \mathcal{I}_1^{-1} \lambda_1 - \mathcal{I}_1^{-1} \mathcal{I}_1 \mathcal{I}_2.1 \mathcal{I}_2.1 \mathcal{I}_2.1 \\
\theta_{2}^{k+1} - \theta_{2}^{k} &= \mathcal{F}_2(\theta_{1}^{k}, \theta_{2}^{k}) = \mathcal{I}_2.1 \mathcal{I}_2.1 \mathcal{I}_2.1 \mathcal{I}_2.1 .
\end{align*}
\]

We may describe (1.32) as parallel iterations for \( \theta_1 \) and \( \theta_2 \).

An attempt to accelerate the convergence of (1.32) is to apply the (nonlinear) Gauss-Seidel principle that information be used as soon as it is available. Using the already updated \( \theta_1 \) to update \( \theta_2 \) gives

\[
\begin{align*}
\theta_{1}^{k+1} - \theta_{1}^{k} &= \mathcal{F}_1(\theta_{1}^{k}, \theta_{2}^{k}) \\
\theta_{2}^{k+1} - \theta_{2}^{k} &= \mathcal{F}_2(\theta_{1}^{k+1}, \theta_{2}^{k})
\end{align*}
\]

(1.33)
which we call alternate iterations. It is worth noting the special case with \( I_{12} = 0 \), for which the alternate iterations are

\[
\begin{align*}
\theta_{1}^{k+1} - \theta_{1}^{k} &= I_{1}^{-1} \lambda_{1}^{k+1} (\theta_{1}^{k}, \theta_{2}^{k}) \\
\theta_{2}^{k+1} - \theta_{2}^{k} &= I_{2}^{-1} \lambda_{2}^{k+1} (\theta_{1}^{k+1}, \theta_{2}^{k}) 
\end{align*}
\]

(1.34)

We consider two other algorithms which attempt to further separate the iterations for \( \theta_{1} \) and \( \theta_{2} \). Let \( \hat{\theta}_{1}(\theta_{2}) \) be the value of \( \theta_{1} \) which maximizes \( \mathcal{L} \) with \( \theta_{2} \) fixed. We define nested iterations to be the process

\[
\begin{align*}
\hat{\theta}_{1}^{k+1} &= \hat{\theta}_{1}(\theta_{2}^{k}) \\
\theta_{1}^{k+1} &= \hat{\theta}_{1}(\theta_{2}^{k+1}), \\
\theta_{2}^{k+1} &= I_{2}^{-1} \lambda_{2}^{k+1} (\hat{\theta}_{1}(\theta_{2}^{k}), \theta_{2}^{k}) 
\end{align*}
\]

(1.35)

Note that for \( \hat{\theta}_{1}^{k+1} \) so defined, \( \lambda_{2,1} = \lambda_{2} \) since \( \lambda_{1} = 0 \). Compare with (1.26). If \( \hat{\theta}_{1}(\theta_{2}) \) is available in closed form, the \( \theta_{1} \) are often said to be separable.

Similarly let \( \hat{\theta}_{2}(\theta_{1}) \) maximize \( \mathcal{L} \) for fixed \( \theta_{1} \). We define coupled iterations to be the process

\[
\begin{align*}
\hat{\theta}_{2}^{k+1} &= \hat{\theta}_{2}(\theta_{1}^{k}) \\
\theta_{2}^{k+1} &= \hat{\theta}_{2}(\theta_{1}^{k+1}), \\
\theta_{1}^{k+1} &= \hat{\theta}_{1}(\hat{\theta}_{2}(\theta_{1}^{k}), \theta_{2}^{k}) 
\end{align*}
\]

(1.36)

Both nested and coupled iterations can be viewed as a reduction in the dimension of the fitting problem, since we can summarize (1.35) as

\[
\begin{align*}
\theta_{2}^{k+1} &= \theta_{2}^{k} + I_{2,1}^{-1} \lambda_{2,1}^{k+1} (\hat{\theta}_{1}(\theta_{2}^{k}), \theta_{2}^{k}) 
\end{align*}
\]

(1.37)

and (1.36) as

\[
\begin{align*}
\hat{\theta}_{2}^{k+1} &= \hat{\theta}_{2}(\hat{\theta}_{1}(\theta_{2}^{k})) 
\end{align*}
\]

(1.38)
without involving \( \theta_1 \).

Coupled iterations consist of iterating each equation of (1.34) to convergence before alternating to the other equation. If \( \theta_1 \) and \( \theta_2 \) are orthogonal then the process is a variation of alternate iterations. We can relate nested to alternate iterations through the transformation (1.22). The alternate iteration for the derived orthogonal parameters is

\[
\begin{align*}
\hat{\theta}_{1.2}^{k+1} - \hat{\theta}_{1.2}^{k} &= I^{-1}_{1.2} \hat{k}_{1.2}(\hat{\theta}_{1.2}^{k}, \hat{\theta}_{2}^{k}) \\
\hat{\theta}_{2}^{k+1} - \hat{\theta}_{2}^{k} &= I^{-1}_{2.1} \hat{k}_{2.1}(\hat{\theta}_{1.2}^{k+1}, \hat{\theta}_{2}^{k}).
\end{align*}
\]

(1.39)

Nested iterations consist of iterating the first equation of (1.39) to convergence before alternating to the second equation. If it happens that \( \hat{\theta}_{1.2}(\theta_2) = \theta_1 + I^{-1}_{1.2} \hat{k}_{1.2}(\theta_1, \theta_2) \) for all \( \theta_2 \), as is the case for linear parameters in least squares, then nested iterations for \( \theta_1 \) and \( \theta_2 \) are exactly equivalent to alternate iterations for \( \theta_{1.2} \) and \( \theta_2 \).

Another useful description of nested iterations, although one which characterizes \( I_{2.1} \) as only one of several possible approximations to the Hessian, is the following. Minimize with respect to \( \theta_2 \) the reduced objective function

\[
\psi(\theta_2) = \ell(\hat{\theta}_{1.2}(\theta_2), \theta_2)
\]

which is not now a likelihood function. We find, as did Richards (1961), that

\[
\ddot{\psi} = \ddot{k}_{2.1}(\hat{\theta}_{1.2}(\theta_2), \theta_2)
\]

and

\[
\dddot{\psi} = \dddot{k}_{2.1}(\hat{\theta}_{1.2}(\theta_2), \theta_2)
\]
where $\hat{\theta}_2.1 = \hat{\theta}_2 - \hat{\theta}_2\hat{\theta}_1^{-1}\hat{\theta}_1$. We apply Newton-Raphson to minimize $\psi$ and approximate $-\hat{\theta}_2.1$ with $I_{2.1}$.

Coupled iterations may be identified as a nonlinear Gauss-Seidel iteration to solve the partitioned $\hat{\lambda} = 0$ (Ortega and Rheinboldt, 1970). We will refer to the calculation of $\hat{\theta}_1(\theta_2)$ and $\hat{\theta}_2(\theta_1)$ as fitting submodels corresponding to the subvectors $\theta_1$ and $\theta_2$ respectively. The terminology is justified by observing that if one subvector is fixed, $\ell(\theta_1, \theta_2)$ may be considered a log-likelihood function of the other subvector alone, possibly in terms of derived data calculated from the original data and the fixed parameters. If $\theta_1$ and $\theta_2$ are orthogonal, then the scoring iteration for $\theta$ may be interpreted as comprising an iteration for $\theta_1$ and an iteration for $\theta_2$, each in their own submodels.

Note the similarity of (1.27) and (1.35). The score test (1.26) for composite hypotheses is equivalent to the Wald test (1.28) but with one nested iteration $\hat{\theta}_2.0 + I_{2.1}^{-1} \hat{\theta}_2(\hat{\theta}_1(\theta_2.0), \theta_2.0)$ in place of $\hat{\theta}_2.2$.

1.12 Convergence Rates of Partitioned Algorithms

1.12.1 Nested iterations

We show that nested iterations have a convergence factor less than or equal to that of the full scoring iteration. Differentiating (1.37) at $\hat{\theta}_2$, and using $\frac{d}{d\theta_2} \hat{\theta}_1(\theta_2) = \hat{\theta}_1^{-1} \hat{\theta}_1$, gives the convergence matrix for nested iterations

$$G = I + I_{2.1}^{-1} \hat{\theta}_2.1.$$

We show that the spectral radius of (1.40) is less than that of (1.14) by showing that the extreme eigenvalues of $I_{2.1}^{-1} \hat{\theta}_2.1$ are bounded by those of $I^{-1} \hat{\theta}_1$. 

Let \( P \) be the \( p \times p \) matrix \((0 \ I)^T\). We observe that
\[ I_{2.1}^{-1} = P^T I^{-1} P \quad \text{and} \quad x_{2.1}^{-1} = P x_{1}^{-1} P. \]
Therefore
\[ I_{2.1}^{-1} x_{2.1} = P^T I^{-1} P (P x_{1}^{-1} P)^{-1}, \]
which has Rayleigh quotient
\[ \frac{z^T P^T I^{-1} P z}{z^T P x_{1}^{-1} P z} \]
for \( z \in \mathbb{R}^p \). The extrema of (1.41) are equivalent to constrained extrema of
\[ \frac{v^T I^{-1} v}{v^T x_{1}^{-1} v} \]
or of
\[ \frac{v^T x_{1} v}{v^T I v} \]
over \( v \in \mathbb{R}^p \), and hence are bounded by the unconstrained extrema.
Observing that (1.42) is the Rayleigh quotient of \( I^{-1} x_{1} \) completes the demonstration.

1.12.2 Coupled iterations

The convergence matrix for coupled iterations emerges, from differentiating (1.38) at \( \hat{\theta} \), as
\[ G = \begin{pmatrix} x_{2.1}^{-1} & x_{12}^{-1} \\ x_{21}^{-1} & x_{12}^{-1} \end{pmatrix} \]
since \( \frac{d}{d \theta_1} \hat{\theta}_2(\theta_1) = -x_{21}^{-1} x_{21}^{-1} \) and \( \frac{d}{d \theta_2} \hat{\theta}_1(\theta_2) = -x_{12}^{-1} x_{12}^{-1} \). If \( x_{1} \) is negative definite, as it will be at a maximum of \( x_{1} \), then the eigenvalues of \( G \) all lie between 0 and 1; they may be recognized as the canonical correlations (e.g. Rao, 1973, §8f.1) calculated from the partitioned covariance matrix
\[
\begin{pmatrix}
-x_{1} & -x_{12} \\
-x_{21} & -x_{2} 
\end{pmatrix}.
\]
This shows that any maxima is a point of attraction for coupled iterations; furthermore the convergence is monotonic.

On the other hand, for $G$ to tend to zero for increasing sample sizes, it is necessary that $I_{21} = \mathbb{E}(\tilde{Z}_{21}) = 0$, that is that $\theta_1$ and $\theta_2$ be orthogonal. If this is not the case, then coupled iterations will be the slowest of the algorithms considered in §1.11 for sufficiently large samples.

1.13 Separable Regression

The most common application of nested iterations is to nonlinear regression. Consider the least squares problem with parameters $\hat{\beta}_T = (\hat{\alpha}, \hat{\beta})$ and with

$$\phi(\alpha, \beta) = (y - \mu)^T(y - \mu).$$

Suppose the mean $\mu$ has the functional form

$$\mu_i = \sum_j x_{ij}(\beta)\alpha_j \quad i = 1, \ldots, n$$

with each $x_{ij}$ a smooth function of $\beta$. The $\alpha_j$ are linear parameters and the $\beta_j$ nonlinear. Gathering the $x_{ij}$ into the matrix function $X$ allows us to write

$$\mu = X(\beta)\alpha..$$

We assume $X$ to be of the full rank, at least in a neighbourhood of the true value $\beta_0$.

The parameters $\alpha_j$ are separable since

$$\hat{\alpha}(\beta) = (X^T X)^{-1} X^T \beta.$$
with \( P_X = X(X^TX)^{-1}X^T \) the projection onto \( R(X) \). Also
\[
I_{2.1}(\hat{\gamma}(\hat{\beta}), \hat{\beta}) = \mu^T_B \mu_B - \mu^T_B (X^TX)^{-1}X^T \mu_B
\]
(1.47)
\[
= \mu^T_B (I - P_X) \mu_B
\]
where \( \mu_B \) is the vector of partial derivatives \( \frac{\partial \mu_i}{\partial \beta_j} \). So
application of nested iterations to the least squares problem gives
\[
\hat{\beta}^{k+1} - \hat{\beta}^k = (\mu_B^T (I - P_X) \mu_B)^{-1} \mu_B^T (\gamma - \mu) .
\]
(1.48)

Separating the linear parameters was suggested by Richards (1961) in a maximum likelihood setting, and by Ross (1970) and Lawton and Sylvestre (1971) for regression. Richards suggested Newton-Raphson to minimize \( \psi(\hat{\beta}) \), while Lawton and Sylvestre suggested finite difference methods. Golub and Pereyra (1973, 1974) applied the formal Gauss-Newton algorithm (1.8) to \( \psi(\hat{\beta}) \) with \( f = (I - P_X)\gamma \), and named the result the \textit{variable projection} algorithm. Kaufman (1975) derived our nested iteration using differentiation of orthogonal matrices. Ruhe and Wedin (1980) showed that the variable projection algorithm and nested iterations have similar convergence factors, and both factors are bounded by that of Gauss-Newton on the unseparated problem. In fact, the eigenvalues of the convergence matrix for the variable projection algorithm are the normal curvatures of the solution locus (multiplied by \( \hat{\gamma} \)), restricted to the sublocus determined by \( \hat{\gamma} = \hat{\gamma}(\hat{\beta}) \).

Each nested iteration requires the same amount of computation as does an iteration of unseparated Gauss-Newton, as can be seen from our derivation in §1.11. Golub and Pereyra (1973) and Ruhe and Wedin (1980) both found that iterations of the variable projection algorithm were more expensive.
1.14 Tests in Submodels

Our aim in this section is to make an observation about orthogonal parameters and to develop the concept of a submodel a little further. Suppose that $\theta$ is partitioned into orthogonal subvectors $\beta$ and $\gamma$, and that these in turn are partitioned into $\beta_1$, $\beta_2$, and $\gamma_1$, $\gamma_2$. The dimensions are $p_1$, $p_2$ and $q_1$, $q_2$ say. Suppose that we wish to test $\beta_2 = \beta_{20}$ and $\gamma_2 = \gamma_{20}$. Then the score test statistic of this combined hypothesis simply consists of the statistics for the hypothesis $\beta_2 = \beta_{20}$ and $\gamma_2 = \gamma_{20}$ tested separately. The same is true of the Wald test, but not of the likelihood ratio test.

Let conformal partitions of $\lambda$ and $I$ be

$$
\lambda = \begin{pmatrix}
\lambda_{\beta_1} \\
\lambda_{\beta_2} \\
\lambda_{\gamma_1} \\
\lambda_{\gamma_2}
\end{pmatrix}
$$

and

$$
I = \begin{pmatrix}
I_{\beta_1} & I_{\beta_{12}} & 0 \\
I_{\beta_{12}} & I_{\beta_2} & 0 \\
0 & I_{\gamma_1} & I_{\gamma_{12}} \\
0 & I_{\gamma_{21}} & I_{\gamma_2}
\end{pmatrix}.
$$

We find the score test (1.26) of $\beta_2 = \beta_{20}$ and $\gamma_2 = \gamma_{20}$ is

$$
\begin{pmatrix}
I_{\beta_2}^{-1/2} \lambda_{\beta_2} \\
I_{\gamma_2}^{-1/2} \lambda_{\gamma_2}
\end{pmatrix} \sim N(0, I)
$$

(1.49)
in an obvious notation. Here \( I_{\hat{\beta}_{2.1}}^{-1} \hat{\mathbf{y}}_{2} \) is the score test of \( \hat{\beta}_2 = \hat{\beta}_{20} \) in the \( \hat{\beta} \)-submodel with \( \gamma \) fixed at \((\hat{\gamma}_1(\gamma_{20}), \gamma_{20})\).

Similarly \( I_{\hat{\gamma}_{2.1}}^{-1} \hat{\beta}_2 \) is the test of \( \hat{\gamma}_2 = \gamma_{20} \) in the \( \gamma \)-submodel with \( \hat{\beta} \) fixed at \((\hat{\beta}_1(\beta_{20}), \beta_{20})\). In its \( \chi^2 \) form

\[
(1.50) \quad \hat{\mathbf{y}}_{2}^T I_{\hat{\beta}_{2.1}}^{-1} \hat{\beta}_2 + \hat{\mathbf{y}}_{2}^T I_{\hat{\gamma}_{2.1}}^{-1} \hat{\gamma}_2 \sim \chi^2_{p+q_2}
\]

the score test emerges as the sum of the statistics from the submodels.

The same observation is true for the Wald test, with

\[
(1.51) \quad \begin{bmatrix} I_{\hat{\beta}_{2.1}}^{-1} (\hat{\beta}_2 - \beta_{20}) \\ I_{\hat{\gamma}_{2.1}}^{-1} (\hat{\gamma}_2 - \gamma_{20}) \end{bmatrix} \sim N(0,I)
\]

and

\[
(1.52) \quad (\hat{\beta}_2 - \beta_{20})^T I_{\hat{\beta}_{2.1}}^{-1} (\hat{\beta}_2 - \beta_{20}) + (\hat{\gamma}_2 - \gamma_{20})^T I_{\hat{\gamma}_{2.1}}^{-1} (\hat{\gamma}_2 - \gamma_{20}) \sim \chi^2_{p+q_2}.
\]

Here (1.51) consists of the Wald test of \( \hat{\beta} = \beta_{20} \) in the \( \hat{\beta} \)-submodel with \( \gamma \) fixed at \( \hat{\gamma} \), and the Wald test of \( \hat{\gamma}_2 = \gamma_{20} \) in the \( \gamma \)-submodel with \( \hat{\beta} \) fixed at \( \hat{\beta} \).

The likelihood ratio test of the combined hypothesis cannot be expressed exactly as the sum of statistics from submodels. This is a consequence of the fact that both restricted and unrestricted estimates of the nuisance parameters are required. An attempted decomposition would take the form of (1.30): the first term is a test in a submodel, but the second is not.

1.15 Asymptotic Stability of the Gauss-Newton Iteration

In this section we outline a proof of the asymptotic stability of Gauss-Newton algorithm, which is essentially that of Jennrich (1969).
We do so in some detail because it foreshadows a proof of the
stability of the Prony algorithm in chapter 7 under very similar
assumptions.

Define the matrix function of $\theta$

$$G_n(\theta) = (\mu(\theta)^T \mu(\theta))^{-1} \mu(\theta)^T (y - y(\theta)) .$$

At $\hat{\theta}$, $G_n$ is the convergence matrix (1.15) of the Gauss-Newton
algorithm. In summary, we prove that $G_n(\hat{\theta})$ tends to zero by proving
that $n^{-1} \mu(\theta)^T \mu(\theta)$ tends to a nonsingular matrix while
$n^{-1} \mu(\theta)^T (y - y(\theta))$ tends to zero. Before we can be more precise we
need to be more specific about our assumptions. Assume that

(a) Error structure. The $y_i - y(\theta_0)$ are independently and
identically distributed with mean zero and variance $\sigma^2$; $\theta_0$ is an
interior point of the compact set $\Theta \subseteq \mathbb{R}^D$.

(b) Unique minimum. The function

$$Q(\theta) = \lim_{n \to \infty} \frac{1}{n} (\mu(\theta) - y(\theta))^T (\mu(\theta) - y(\theta))$$

has a unique minimum over $\Theta$ at $\theta_0$.

(c) Smoothness of $\mu$. All limits of the form $\lim \frac{1}{n} \frac{\partial^2 \mu}{\partial \theta_i \partial \theta_j}$ exist and are continuous on $\Theta$.

(d) Nonsingular information matrix. The matrix $A$ defined by

$$A_{ij} = \lim_{n \to \infty} \frac{1}{n} \frac{\partial^2 \mu}{\partial \theta_i \partial \theta_j}$$

is nonsingular at $\theta_0$.

We need the following form of the Law of Large Numbers.

Theorem 1. If $\{f_i\}$ is a sequence of continuous functions on $\Theta$
such that
\[ \frac{1}{n} \sum_{i=1}^{n} f_i(\theta_1) f_i(\theta_2) \]
converges uniformly on \( \Theta \times \Theta \), then
\[ \frac{1}{n} \sum_{i=1}^{n} f_i(\theta)(y_i - \mu_i(\theta_0)) \xrightarrow{a.s.} 0 \]
uniformly on \( \Theta \).


We will accept the consistency of \( \hat{\theta} \) as having been established (see Jennrich (1969)), and are now in a position to prove

Theorem 2. Under conditions (a) to (d),
\[ G_n(\hat{\theta}) \xrightarrow{a.s.} 0. \]

Proof. We must establish that \( G_n \) exists (i.e. \( \mu^T \mu \) is nonsingular) and converges, uniformly in some neighbourhood \( S \) of \( \Theta_0 \), to a continuous function \( G \) which is zero at \( \Theta_0 \). Having done so we can write
\[ \| G_n(\theta) \| \leq \| G_n(\theta) - G_n(\Theta_0) \| + \| G_n(\Theta_0) - G(\Theta_0) \| \]
where \( \| \cdot \| \) is any matrix norm. We can make \( G_n(\theta) - G_n(\Theta_0) \) small by choosing \( \theta \) within a suitably small neighbourhood \( S_1 \subseteq S \) of \( \Theta_0 \), and \( G_n(\Theta_0) - G(\Theta_0) \) small by choosing \( n \) large. Finally, using consistency of \( \hat{\theta} \), we can choose \( n \) large to ensure \( \hat{\theta} \in S_1 \) a.s. thus proving the theorem.

The existence and convergence of \( G_n \) is established as follows. Since \( A \) is a continuous and nonsingular at \( \Theta_0 \), it is nonsingular
on a compact neighbourhood $S$ of $\theta_0$. Since $n^{-1} \mu^T \mu + A$ uniformly, $\mu^T \mu$ is nonsingular in $S$ for $n$ sufficiently large. We can write then, for $\theta \in S$,

$$G_n(\theta) = \left( \frac{1}{n} \mu(\theta)^T \mu(\theta) \right)^{-1} \frac{1}{n} \mu(\theta)^T \{ (y - \mu(\theta_0)) + (\mu(\theta_0) - \mu(\theta)) \}.$$ 

By the Law of Large Numbers $n^{-1} \mu(\theta)^T (y - \mu(\theta_0)) \xrightarrow{a.s.} 0$. Let

$$F(\theta) = \lim_{n \to \infty} \frac{1}{n} \mu(\theta)^T (\mu(\theta_0) - \mu(\theta)).$$

We find that $G_n(\theta) \xrightarrow{a.s.} A(\theta)^{-1} F(\theta)$ uniformly for $\theta \in S$, and observe that $F$ is zero at $\theta_0$.

Corollary. *The Gauss-Newton iteration is asymptotically stable.*

**Remark on asymptotic arguments**

The operational content of an asymptotic argument such as the above is that it leads, at least implicitly, to an expansion for the quantity of interest, in this case $\rho(G_n(\hat{\theta}))$, in terms of increasing negative powers of $n$. This expansion then provides an approximation which is applied in finite samples.

Assumptions (a) to (d) above ensure that the information matrix $\mu^T \mu$ is of order $n$. This is a stronger growth condition than necessary. Wu (1981) has shown that, for consistency of $\hat{\theta}$, a necessary condition is that

$$Q_n(\theta) = (\mu(\theta) - \mu(\theta_0))^T (\mu(\theta) - \mu(\theta_0)) \to \infty$$

for all $\theta \neq \theta_0$. Condition (1.53) is also sufficient when combined with other assumptions which basically ensure that the minimum eigenvalue of $\mu^T \mu$ grows at a greater rate than the square root of the maximum eigenvalue. The value of Wu's results is that they may
provide more information about constants appearing in the above mentioned expansion.

It is also assumed in the above proof that \( \theta \) is restricted to a compact set \( \Theta \). In many specific examples it is possible to prove that the unrestricted least squares estimate must eventually belong to a compact set containing \( \Theta_0 \), thus making the prior assumption unnecessary. This is an important point, because if compactness of \( \Theta \) was critical, and \( \Theta \) had to be chosen very large indeed in a specific example, then there would be no reason to expect that constants appearing in expansions for \( \rho(G_n(\hat{\theta})) \) would not also be very large.
PART I
COUPLED ITERATIONS AND VARYING PARAMETER MODELS

PREAMBLE

Chapters 2, 3 and 4 deal with a number of varying parameter models. By this we mean that each observation $y_i$ is sampled from a distribution known up to parameters $\theta_1, \ldots, \theta_p$, and that one or more of the parameters vary as a function of $i$. This is, of course, extremely general, but we have in mind the case for which $y_i$ is sampled at time $t_i$ and the distribution drifts gradually with time.

There is a vast time series and econometrics literature on varying parameter models, for example Brown, Durbin and Evans (1975), Raj and Ullah (1981), Nicholls and Quinn (1982), Chow (1983), and Nicholls and Pagan (1985). West, Harrison and Migon (1985) use a Bayesian state space formulation to fit varying parameter generalized linear models. Several factors distinguish the material in the chapters which follow from most of the above literature. We will give most attention to independent observations, and will usually assume that the varying parameter itself depends in a known way on other parameters, which are to be estimated. In fact we will usually assume that the parameter variation is linear on an appropriate scale, for example $g(\theta_{p,i}) = \beta_1 + \beta_2 t_i$ with $g$ a known monotonic function. (It introduces no new problems to go at once to the more general $g(\theta_{p,i}) = x_i^T \beta$ with $x_i$ any vector of covariates.) We give special attention to varying parameters which relate to the variance or shape.
of the distribution, and attempt to treat them in a way analogous to that of parameters which determine the mean.

The main examples of specific varying parameter models which we consider are the following.

(a) Generalized linear models with varying dispersion parameters. That is, we consider distributions with densities of the form

$$\exp\{\phi^{-1}(y\nu - b(\nu)) + c(y, \phi)\}$$

and allow both $\nu$ and $\phi$ to vary.

(b) Normal autoregressive models. The varying distribution is the conditional distribution of $y_i$ given $y_1, \ldots, y_{i-1}$, which is defined by the unconditional expectation $\mu$, the autoregressive coefficients $\beta_1, \ldots, \beta_p$, and the unconditional variance $\sigma^2$. We allow $\mu$, the $\beta_j$, and $\sigma^2$ all to vary.

(c) Multinormal models. The varying distribution is itself $p$-variate normal with mean $\mu$ and covariance matrix $\Sigma$. We parametrize in terms of the elements of $L^{-1}$ and $D$, where $L$ is lower triangular and $D$ diagonal and $\Sigma = LDL^T$, and allow all parameters to vary.

The models have in common that their parameters may be divided into orthogonal subsets possessing relatively simple submodels. Coupled iterations may therefore be usefully applied to obtain numerical parameter estimates, and the associated submodels provide a convenient framework for inference and interpretation.
CHAPTER 2

GENERALIZED LINEAR MODELS WITH VARYING DISPERSION PARAMETERS

2.1 Introduction

To apply classical linear regression it is necessary to find some transformation of the data which will simultaneously combine approximate normality, additivity of systematic effects, and constant variance. Generalized linear models substantially relax this requirement by providing a choice of distributions and by allowing additivity of systematic effects to hold on a transformed scale, the choice of scale being made independently of the choice of distribution. On the other hand, once these choices have been made, the variance is then specified as a function of the mean up to a multiplicative constant. In this chapter we free the variance from this dependence by allowing the multiplicative constant (the dispersion parameter) to depend on covariates and unknown parameters in the same way as does the mean.

2.2 Mean and Variance Structure

Let \( y \) be a data vector of independent observations. Write \( \mathbb{E}(y) = \mu \) and \( \mathbb{D}(y) = \langle \sigma^2 \rangle \). Quasi-likelihoods and generalized linear models assume that

\[
(2.1) \quad g(\mu) = X\beta
\]

with \( g \) a scalar, differentiable, monotonic function, and \( X \) a matrix of covariates, and that
with $\psi$ a scalar non-negative function, and $w$ a vector of known weights. (Recall that scalar functions act component-wise on vectors, and that juxtaposition of column vectors denotes component-wise products.) Here $\phi$ is the dispersion parameter and $v$ is the variance function. We generalize to

$$\sigma^2 = \phi v(\mu)$$

with $\phi$ a vector of dispersion parameters, and assume that

$$h(\phi) = Z\gamma$$

with $h$ a differentiable monotonic function.

The possibility of applying (2.2) to generalized linear models was mentioned by Pregibon (1984), who noted that it appeared to be unresearched. In the context of normal linear regression, several particular choices for $h$ have been suggested in the applied statistics literature. For example Park (1966), Harvey (1976), and Cook and Weisberg (1983) used $\log \sigma^2 = Z\gamma$. Glejser (1969) suggested a linear model for $\sigma$ or for $\sigma^2$ itself. Breusch and Pagan (1979) used the general $h(\sigma^2) = Z\gamma$, but were concerned only with testing for heteroscedasticity and not with estimation of $\gamma$.

2.3 Exponential Family Likelihoods

Generalized linear models assume a log-likelihood function of the form

$$\ell = \Sigma_i \{ w_i \psi^{-1}(y_i, v_i) - b(v_i) + c(y_i, w_i \psi^{-1}) \}.$$  

It follows that $\mathbb{E}(y_i) = \mu_i = \hat{b}(v_i)$ and $\mathbb{D}(y_i) = w_i \phi^{-1} v(\mu_i)$ with $v(\mu_i) = \hat{b}(v_i)$. If $\phi$ is fixed then (2.3) defines an exponential
family with canonical parameters \( \nu_i \). With \( \phi \) a parameter, (2.3) defines an exponential family only if \( c \) is of the form

\[
c(y, w^{\phi^{-1}}) = w^{\phi^{-1}}a(y) - s(w^{\phi^{-1}}) + f(y)
\]

for some functions \( a, s \) and \( f \). This is so for the normal, inverse Gaussian and gamma distributions, which are the two parameter distributions usually used in generalized linear models.

We will assume (2.4), and generalize to unequal dispersion parameters. The log-likelihood then becomes

\[
\ell = \sum_i \left( w_i^{\phi_i^{-1}}(y_i \nu_i - b(\nu_i) + a(\nu_i)) - s(w_i^{\phi_i^{-1}}) + f(\nu_i) \right).
\]

Here the term \( f(\nu_i) \) can be removed by an appropriate transformation of \( \nu \).

2.4 The Mean Submodel

Since the dispersion parameter is no longer constant, it does not cancel out of the scoring iteration for \( \beta \) as in §1.7. Instead we use coupled iterations as described in §1.11. Firstly we note that

\[
\frac{\partial^2 \ell}{\partial \phi_i \partial \nu_i} = -w_i^{\phi_i^{-2}}(y_i - \nu_i)
\]

which has expectation zero. Therefore \( \nu \) and \( \phi \), and hence \( \beta \) and \( \nu \), are orthogonal.

For fixed \( \phi \), (2.5) is a log-likelihood for a generalized linear model with weights \( w_i^{\phi_i^{-1}} \) and dispersion parameter \( 1 \). Combining with (2.1) completes the generalized linear model formulation. We call this the mean submodel.
2.5 The Deviance Residuals

Let

\[ Q(y, \mu) = yv - b(v) + a(y) . \]

As a function of \( \mu \), \( Q \) is maximized by \( \mu = y \), and is generally finite. Therefore each

\[ d_i(y_i, \mu_i) = 2w_i(Q(y_i, y_i) - Q(y_i, \mu_i)) \]

takes values in the half line \([0, \infty)\) as \( \mu_i \) varies. Let

\[ \lambda = \sum_i \phi_i \frac{-1}{\hat{d}_i} \]

with \( \hat{d}_i = d_i(y_i, \hat{\mu}_i) \). Thus \( \lambda \) is the likelihood ratio test, in the mean submodel, that \( \mu \) is described by (2.1) against the alternative that it is unrestricted. Hence it may be interpreted as a measure of the goodness of fit of \( \hat{\mu} \) to \( y \), and has asymptotic distribution \( \chi^2_{n-p} \). We carry the usual terminology of generalized models into this more general setting by calling \( \lambda \) the deviance and the

\[ \text{sign}(y_i - \hat{\mu}_i) \hat{d}_i \]

the deviance residuals (McCullagh and Nelder, 1983).

2.6 The Dispersion Submodel

In terms of the \( d_i \), the log-likelihood (2.5) may be written

\[ (2.6) \quad \ell = \frac{-1}{2} \sum_i \{ \phi_i^{-1} d_i + 2s(w_i \phi_i^{-1}) + \text{(function of } y_i) \} . \]

For fixed \( \mu \), (2.6) defines an exponential family with data \( d_i \) and canonical parameters \( -\phi_i^{-1} \). Together with (2.2), we call this the dispersion submodel. We see that

\[ \mathbb{E}(d_i) = \sigma_i^2 = -2w_i s(w_i \phi_i^{-1}) \]

and

\[ \mathbb{D}(d_i) = 2\sigma_i^2 = w_i^2 s(w_i \phi_i^{-1}) . \]
We define the $\sigma_i^2$ to be the dispersion mean parameters and the $u_i$ to be the dispersion variance functions. The dispersion submodel itself has a dispersion parameter, which is 2.

Note that $h$ is defined by (2.2) to be a function of the $\phi_i$, which are not necessarily equal to the $\sigma_i^2$. But if we define $f$ to be the implicit dispersion link satisfying

$$f(\sigma^2) = z\gamma$$

then we must have

$$\dot{f}_i = \phi_i^2 u_i^{-1} \cdot$$

Using $\sigma^2$, $u$ and $\dot{f}$, the scoring iteration for $\gamma$ may be written in the form (1.12) as for a generalized linear model, the difference being that $\sigma_i^2$, $u_i$ and $\dot{f}_i$ depend on $w_i$ as well as on $(Z\gamma)_i$.

The dispersion model is exactly a generalized linear model whenever the dependence of $\sigma_i^2$ and $u_i$ on $w_i$ is removed. This occurs whenever the prior weights $w_i$ are equal or the function $s$ is logarithmic. The latter is true for the normal and inverse Gaussian distributions.

2.7 The Normal Distribution

2.7.1 The mean submodel

The normal distribution has log-likelihood

$$\ell = -\frac{1}{2} \Sigma_i w_i \sigma_i^{-2} (y_i - \mu_i)^2 - \frac{1}{2} \Sigma_i \log(w_i \sigma_i^{-2}) - \frac{n}{2} \log 2\pi$$

where $\mu_i$ is the mean and $w_i \sigma_i^{-2}$ the variance of $y_i$. In the notation of §2.3 we have $\nu_i = \mu_i$, $b(\nu) = \nu^2$ and $v(\mu) = 1$. The mean submodel is a generalized linear model with means $\mu_i$, weights $w_i \sigma_i^{-2}$, and variance function identically 1.
2.7.2 The dispersion submodel

The deviance components are \( d_i = w_i(y_i - \mu_i)^2 \), which are distributed as \( \sigma_i^2 \chi_1^2 \) if the \( \mu_i \) are known. (Recall that \( \chi_1^2 = 2G(\frac{1}{2}) \), where \( G(\frac{1}{2}) \) represents a Gamma random variable with shape parameter \( \frac{1}{2} \).) In the notation of §2.6, \( 2s(w^{-1}) = \log(w^{-1}) \), the parameters \( \sigma_i^2 \) and \( \phi_i \) coincide, as do the functions \( f \) and \( h \), and \( w(\sigma^2) = \sigma^4 \).

2.8 The Inverse Gaussian Distribution

2.8.1 The mean submodel

The inverse Gaussian log-likelihood is

\[
\ell = -\frac{1}{2} \sum_i w_i \lambda_i \frac{(y_i - \mu_i)^2}{\mu_i^2 \nu_i} + \frac{1}{2} \sum_i \log(w_i \lambda_i) - \frac{1}{2} \sum_i \log(2\pi \nu_i^3)
\]

with the \( y_i, \mu_i \) and \( \lambda_i \) all positive. We have \( \nu_i = -\frac{1}{2} \mu_i^{-2} \), \( b(\nu) = -(-2\nu)^{\frac{1}{2}} \), \( \hat{b}(\nu) = \nu \), and \( v(\mu) = \hat{b}(\nu) = \mu^3 \). The mean submodel is a generalized linear model with means \( \mu_i \) weights \( w_i \lambda_i \), and variance function \( \mu^3 \).

2.8.2 The dispersion submodel

The deviance components are \( d_i = w_i \frac{(y_i - \mu_i)^2}{\mu_i^2 \nu_i} \), which are distributed as \( \lambda_i^{-1} \chi_1^2 \) if the \( \mu_i \) are known. The dispersion submodel is therefore exactly as for the normal distribution with \( \phi_i = \sigma_i^2 = \lambda_i^{-1} \).
2.9 The Gamma Distribution

2.9.1 The mean submodel

The gamma density may be written

\[
\frac{y^{\alpha-1} e^{-y/\beta}}{\beta^\alpha \Gamma(\alpha)}
\]

in terms of the shape parameter \( \alpha \) and scale parameter \( \beta \). (If \( \beta = 1 \) we denote a random variable so distributed by \( G(\alpha) \).) The mean of \( y \) is \( \mu = \alpha \beta \). In terms of \( \mu \) and \( \alpha \), the gamma log-likelihood is

\[
\ell = \sum \left\{ w_i \alpha \left( \frac{y_i}{\mu_i} - \frac{\psi_i}{\mu_i} \right) + w_i \alpha \log(w_i \alpha_i) - \log \Gamma(w_i \alpha_i) - \log y_i \right\}.
\]

We have \( \psi_i = -\mu_i^{-1} \), \( b(\psi) = -\log(-\psi) \), \( \hat{b}(\psi) = \mu \) and \( \nu(\mu) = \hat{b}(\psi) = \mu^2 \).

The mean submodel is a generalized linear model with means \( \mu_i \), weights \( w_i \alpha_i \), and variance function \( \mu_i^2 \).

2.9.2 The dispersion submodel

The deviance components are

\[
d_i = 2 \left( \frac{y_i}{\mu_i} - \log \frac{\mu_i}{\psi_i} \right), \quad \phi_i = \alpha_i^{-1},
\]

and

\[
s(\psi^{-1}) = \log \Gamma(\psi^{-1}) - \psi^{-1} \log(\psi^{-1}).
\]

The dispersion mean parameters and variance functions are given by

\[
\mathbb{E}(d_i) = \sigma_i^2 = 2w_i(\psi(w_i \alpha_i) - \log(w_i \alpha_i) - 1)
\]

and

\[
\mathbb{D}(d_i) = 2w_i(\sigma_i^2) = 4w_i^2(\psi(w_i \alpha_i) - (w_i \alpha_i)^{-1})
\]

where \( \psi \) is the digamma function.

2.10 Likelihood Calculations

We summarize here the likelihood calculations of the previous sections. In the notation of §2.2 and §2.3, the score vector and
and the dispersion parameters \( \phi \) have been absorbed into \( w' = \phi^{-1} \).

In the notation of §2.6, the score vector and information matrix for \( \gamma \) are

\[
\ell_{\gamma} = \mathbf{Z}^T \left( \frac{-1}{2f_u} \right) (\mathbf{d} - \mathbf{g}^2)
\]

and

\[
I_{\gamma} = \mathbf{Z}^T \left( \frac{-1}{2f_u^2} \right) \mathbf{Z}.
\]

The scoring iterations therefore are

\[
\beta^{k+1} = F_1(\beta^k, \gamma)
\]

with

\[
F_1(\beta, \gamma) = \beta + I_{\beta}^{-1} \ell_{\beta}
\]

and

\[
\gamma^{k+1} = F_2(\beta, \gamma^k)
\]

with

\[
F_2(\beta, \gamma) = \gamma + I_{\gamma}^{-1} \ell_{\gamma}.
\]

For completeness we also give the full hessian of the log-likelihood, the components of which are

\[
-\mathbf{H}_{\beta} = I_{\beta} + \mathbf{X}^T \left( \mathbf{w}\mathbf{g}^{-1} \mathbf{g}\mathbf{v}^{-1} \mathbf{v} (\mathbf{y} - \mu) \right) \mathbf{X}
\]

and

\[
-\mathbf{H}_{\beta\gamma} = \mathbf{X}^T \left( \mathbf{w}\mathbf{g}^{-1} \mathbf{v} (\mathbf{y} - \mu) \right) \mathbf{Z}
\]

and

\[
-\mathbf{H}_{\gamma} = I_{\gamma} + \mathbf{Z}^T \left( \mathbf{Z}_{\mathbf{W}} \mathbf{u}^{-1} (\mathbf{y} - \mu) \right) \mathbf{Z}.
\]

Remarks (i) Orthogonality and information

The orthogonality of \( \beta \) and \( \gamma \) gives us some protection against
dilution of information as extra parameters are introduced into the model. For example, if a model with varying $\phi$ is mistakenly fitted to data for which the dispersion is actually constant, the information matrix for $\beta$ is asymptotically equivalent to what it would have been had the correct model been fitted. To first order then, the precision with which $\beta$ is estimated is unaffected.

(ii) Variance stabilizing link functions

If a generalized linear model happens to have a link function which is the variance stabilizing transformation for the distribution concerned, then it happens that $g_v$ is identically 1 (for the dispersion submodel) so that the weight matrix which appears in the scoring iteration does not need to be recalculated at each step. Examples are $g(\mu) = \mu^{-2}$ for the inverse Gaussian distribution and $g(\mu) = \log \mu$ for the gamma distribution. It is interesting to note that in this case the difference $|g(\mu_1) - g(\mu_2)|$, where $\mu_1$ and $\mu_2$ are any two values for the mean, is proportional to the distance between the corresponding probability distributions at $\mu = \mu_1$ and $\mu = \mu_2$, as measured by Rao distance (Atkinson and Mitchell, 1981). (The information matrix defines a local Riemannian metric on the parameter space. The Rao distance between two parameter values is then the integrated arc length of a geodesic connecting them.) For example, the Rao distance between the distributions $\sigma_1^{-1} \alpha^{-1} G(\alpha)$ and $\sigma_2^{-1} \alpha^{-1} G(\alpha)$, and hence between $N(\mu, \sigma_1^2)$ and $N(\mu, \sigma_2^2)$, is proportional to $|\log \sigma_1^2 - \log \sigma_2^2|$. The practical import of this is not at all clear, but it may affect convergence of the scoring iteration.
2.11 Three Algorithms

We suggest the following three algorithms to obtain maximum likelihood estimates for $\beta$ and $\gamma$. The first consists of the scoring iteration to minimize the log-likelihood (2.5) with respect to $\beta$ and $\gamma$ simultaneously. The second and third consist, respectively, of alternate and coupled iterations, as described in §1.11. Recall that $F_1$ and $F_2$ are the scoring iteration functions for $\beta$ and $\gamma$ respectively.

Algorithm 1: scoring

\[
\begin{align*}
  k &:= 0 \\
  \text{repeat} & \quad \\
  & \quad \begin{cases}
    \beta^{k+1} := F_1(\beta^k, \gamma^k) \\
    \gamma^{k+1} := F_2(\beta^k, \gamma^k) \\
    k := k+1
  \end{cases} \\
  \text{until convergence}
\end{align*}
\]

Algorithm 2: alternate iterations

\[
\begin{align*}
  k &:= 0 \\
  \text{repeat} & \quad \\
  & \quad \begin{cases}
    \beta^{k+1} := F_1(\beta^k, \gamma^k) \\
    \gamma^{k+1} := F_2(\beta^{k+1}, \gamma^k) \\
    k := k+1
  \end{cases} \\
  \text{until convergence}
\end{align*}
\]
Algorithm 3: coupled iterations

\[ k := 0 \]

repeat

\[ j := 0 \]

repeat

\[ \beta^{j+1} := F_1(\beta^j, \gamma^k) \]

\[ j := j+1 \]

until convergence in the mean submodel

\[ \beta^{k+1} := \beta^{j+1} \]

\[ j := 0 \]

repeat

\[ \gamma^{j+1} := F_2(\beta^{k+1}, \gamma^j) \]

\[ j := j+1 \]

until convergence in the dispersion submodel

\[ k := k+1 \]

until overall convergence

The rates of convergence for these algorithms were discussed in chapter 1, especially in §1.12. In particular, the convergence factor for Algorithm 3 depends on the components of the hessian \( \hat{\mathcal{H}} \), which were given in §2.7. It is worth noting the special case for which the mean model is simply weighted linear regression, so that

\[ \beta^{k+1} = F_1(\beta^k, \gamma) \] converges in one iteration. In that case Algorithm 2 is equivalent to nested iterations (§1.11) with \( \beta \) being separated out, while Algorithm 3 is equivalent to nested iterations with \( \gamma \) separated out. Hence both algorithms are faster than Algorithm 1 in terms of the number of overall iterations required. Since Algorithms 1 and 2 require the same calculations per iteration, Algorithm 2 is then faster in terms of calculations required as well.
2.12 Starting Values

For computing purposes it is convenient to put the scoring iteration for \( \hat{\beta} \) in the form of the weighted linear regression (1.13), and similarly for \( \gamma \). An advantage of the iteration (1.13) is that the right hand side is a function of \( \hat{\beta} \) only through \( \mu \), and sometimes a value for \( \mu \) may be used such that \( g(\mu) \) does not belong to \( \mathcal{R}(X) \). This allows the data itself to be used as a starting value, by putting \( \mu^0 = y \) \((\sigma^2)^0 = \sigma^2 \) in the dispersion submodel). This is equivalent to regressing \( g(y) \) on \( X \) (\( f(w) \) on \( Z \) in the dispersion submodel), which makes especial sense if \( g \) is the variance stabilizing transformation. It is natural to put \( \phi = 1 \) the first time the mean model is fitted.

As an example of using the data as a starting value, consider the case of gamma observations with means \( \sigma^2_i \) and log-link. That is, suppose \( y_i \sim \sigma^2_i \alpha^{-1} G(\alpha) \) for \( i = 1, \ldots, n \), and \( \log \sigma^2 = Z_Y \). Then

\[
\mathbb{E}(\log y) = Z_Y - \log \alpha + \psi(\alpha)
\]

and

\[
\mathbb{D}(\log y) = \psi^2(\alpha) I.
\]

Therefore regressing \( \log y \) on \( Z \) will consistently estimate \( \gamma \) if the constant term is decreased by \( \psi(\alpha) - \log \alpha \). Equivalently, \( \sigma^2 \) will be consistently estimated if it is multiplied by \( \alpha e^{-\psi(\alpha)} \). The efficiency of this estimator is \( \alpha^2 \psi(\alpha)^{-1} \). In the case that \( y_i \sim \sigma^2_i \chi^2_1 \), we have \( \alpha = \frac{1}{\alpha} \), \( \psi(\alpha) - \log \alpha = 1.27036 \), \( \alpha e^{-\psi(\alpha)} = 3.56214 \), and \( \alpha^2 \psi(\alpha)^{-1} = 4\pi^{-2} \).

The above example may be applied to the dispersion submodel if the data is normal or inverse Gaussian. Suppose the dispersion link function \( f = h = \log \), and write \( \mu = G_1(\mu, \sigma^2) \) and \( \sigma^2 = G_2(\mu, \sigma^2) \) for the scoring iterations in terms of \( \mu \) and \( \sigma^2 \). Then the sequence...
1. Fit mean model with $\sigma^2 = \ell$, producing $\mu$

2. $(\sigma^2)^{-1} = G_2(\mu^1, \ell) \times 3.56214$

3. $(\sigma^2)^2 = G_2(\mu^1, (\sigma^2)^1)$

4. $\mu^2 = G_1(\mu^1, (\sigma^2)^2)$

produces efficient estimates for all parameters. In the normal case this is the "three step" estimator of Harvey (1976).

2.13 Power

For there to be a good chance of detecting changes in the dispersion, the components of $\gamma$ must be reasonably large compared with their standard errors. We show through an example that only moderate sample sizes may be required.

Consider again the case of normal or inverse Gaussian data with $f = h = \log$. Suppose that $Z = (\widetilde{Z})$ and that the $z_i$s are equi-spaced, $z_{i+1} - z_i = \delta$ say, and add to zero. The standard error of $\hat{\gamma}_2$ then is

$$(\delta^2 T Z)_{-1/2} = \left(\frac{\delta^2}{12} n(n+1)(2n+1)\right)^{-1/2}.$$ 

Suppose also that $\sigma^2$ doubles over the observed range of the $z_i$, that is $\gamma_2 = (n\delta)^{-1} \log 2$. For $\gamma_2$ to be more than twice standard error with which it is estimated, a sample size of 49 is required.

2.14 Asymptotic Tests

Score and Wald tests of hypotheses may be calculated routinely from the score vectors and information matrices given in §2.10. For likelihood ratio tests it is necessary to distinguish between tests in
the submodels and tests in the overall model. Likelihood ratio tests in the mean submodel can be calculated by differencing values of the deviance corresponding to nested hypotheses. A definition of deviance can be constructed for the dispersion submodel also, directly analogous to that of §2.5 for the mean submodel. Differencing this then results in likelihood ratio tests in the dispersion model. For any model that is fitted the deviances in the mean and dispersion submodels provide goodness of fit tests for the mean and variance structure, respectively, taken one at a time. This will usually be sufficient.

Once again we consider in more detail the case for which the dispersion submodel is a gamma generalized linear model with log link. Let \((y_1^T, y_2^T)^T\) and \((z_1^T, z_2^T)^T\) be conformal partitions of \(y\) and \(z\).

The score test of \(y_2 = 0\) is

\[
2^{1/2} (z_2^T z_2 - z_2^T z_1 (z_1^T z_1)^{-1} z_1^T z_2) - z_2^T (z_2^{-2} d - d) \sim N(0, 1).
\]

If \(z_1 = 1\), so that \(y_2 = 0\) corresponds to constant dispersion, then the \(\chi^2\) version of (2.7) can be interpreted as half the sum of squares due to regression of \(z_2^{-1} d\) on \(z_2\). This interpretation was noted in the normal case by Breusch and Pagan (1979) and Cook and Weisberg (1983). The score test of constant dispersion takes the same form whatever link function is assumed.

In this special case, the overall likelihood ratio test also has a simple form. Suppose we wish to compare the hypotheses \(H_F\) and \(H_R\), with maximum likelihood estimates \((\hat{\beta}_F, \hat{\gamma}_F)\) and \((\hat{\beta}_R, \hat{\gamma}_R)\), \(H_R\) being a restricted version of \(H_F\). From (2.6), twice the difference in the likelihoods is
If \( h \) is logarithmic and \( R(Z) \) contains the constant vector, then \( \hat{\gamma} = 0 \) ensures that

\[
\Sigma_i \hat{\phi}_R^{-1} d_R = \Sigma_i \hat{\phi}_F^{-1} d_F = n
\]

so that (2.8) simplifies to

\[
2 \Sigma_i \{ s(w_i \hat{\phi}_R^{-1}) - s(w_i \hat{\phi}_F^{-1}) \}.
\]

If \( s \) is also logarithmic, then this further simplifies to

\[
(2.9) \quad Z^T Z (\hat{\gamma}_R - \hat{\gamma}_F).
\]

It is an interesting result in itself that (2.9) is non-negative. The simplest of all is the test of constant dispersion, which is

\[
n(\hat{\gamma}_R - \hat{\gamma}_F)
\]

if \( 1 \) is the first column of \( Z \) and the other columns are chosen to add to zero. A similar version was given in the normal case by Harvey (1976).

2.15 Residuals and Influential Observations

In the preceding sections we have shown how the calculations associated with maximum likelihood estimation for our proposed model may be performed separately in the mean and dispersion submodels. This suggests that model checking might also be done in the submodels. For any of the various model checking criteria available in generalized linear models (see for example Pregibon (1981, 1982), McCullagh and Nelder (1983)), we would obtain two diagnostics, one calculated in the mean submodel and one in the dispersion submodel. An issue which
naturally arises is the relationship of the diagnostics to each other and to the overall model.

For example, consider again the scoring iteration (1.13) for generalized linear models arranged as a linear regression. On each iteration, the "hat" matrix \( H = W^T X W X W^T \) projects the vector \( W^T \hat{z} \) onto \( R(X) \). A necessary, and nearly sufficient, condition for the estimates to be consistent and asymptotically normal is that the diagonal elements of \( H \) should tend to zero. Large diagonal elements indicate influential observations (Hoaglin and Welsch, 1978). In our analysis we have two hat matrices, which indicate the influence of observations on the fitted means and dispersions respectively.

McCullagh and Nelder (1983) discuss several ways to define residuals for generalized linear models. Whatever definition is used, we will obtain one residual which measures the difference between \( y_i \) and \( \mu_i \) in the mean submodel, and one which measures the difference between \( d_i \) and \( \sigma_i^2 \) in the dispersion submodels. By way of example, a logical way to check the corrections of the assumed variance structure, would be to inspect plots of residuals from the dispersion submodel.

We devote the remainder of this section to working through two specific criteria, one for an outlier, and one for an observation to be influential with respect to the parameter estimates. These criteria exist in the literature, and are intended to further illustrate the duality of diagnostics in the mean and dispersion submodels.

Consider the hypothesis that observation \( y_j \) is an outlier, that is that \( y_j \) was generated by a process different from that which
generated the remainder of the data. Outliers are commonly modelled
by supposing them to be mean contaminated (Cook and Weisberg, 1982).
Accordingly, as in Pregibon (1982), we modify (2.1) to

\[(2.10) \quad g(y_j) = x_j^\beta + e_j^\alpha\]

where \(e_j\) is the \(j\)th coordinate vector, so that \(y_j\) is an outlier
if \(\alpha \neq 0\). The score test of \(\alpha = 0\) in the mean submodel is

\[(2.11) \quad \sum_j \left( \frac{y_j - \mu_j}{\phi_j \sqrt{z_j^2 (1 - h_j)}} \right) \]

where \(h_j = e_j^T w x (x^T w x)^{-1} x^T w e_j\) and \(w = (g^{-2} - 1)^{1/2}\). The \(t_j\) may
also be interpreted as standardized residuals. (Another definition of
generalized residuals, that of deviance residuals, was given in §2.4.)

An alternative outlier model is to assume that the outlier arises
from a process with an inflated variance (Cook et al., 1982, in the
context of normal linear regression). We modify (2.2) to

\[(2.12) \quad h(\phi) = z^2 + e_j \alpha \quad \alpha \geq 0\]

and again test \(\alpha = 0\). (We have assumed \(h > 0\). If \(h < 0\) then
\(\alpha \leq 0\).) The score test in the dispersion submodel is

\[(2.13) \quad s_j = \frac{d_j - \sigma_j^2}{2 \sqrt{d_j (1 - k_j)}}\]

with \(k_j = e_j^T v z (z^T v z)^{-1} z^T v e_j\) and \(v = (z^2)^{-1}\). The \(s_j\) may
be interpreted as residuals in the dispersion submodel.

The two tests (2.11) and (2.13) will not necessarily select the
same observations as outliers. Indeed the two tests and their
underlying models (2.10) and (2.12) represent different approaches to the treatment of outliers. From the point of view of the mean submodel, fitting (2.10) is equivalent to discarding the jth observation, while fitting (2.12) accommodates it into the model with a decreased weight, namely \( w_j \phi_j^{-1} \). (See Barnett and Lewis, 1978, for general comments on discarding or accommodating outliers.) Using the methods described in this chapter, it is easy to accommodate any set of outliers in this way. For example, if observations \( j_1, \ldots, j_r \) are supposed to be outliers, we might fit
\[
h(\phi) = \gamma_y + (e_{j_1} + \ldots + e_{j_r}) \alpha \quad \alpha \geq 0.
\]

A measure of the influence of the jth observation on \( \hat{\beta} \) can be obtained by comparing the change in \( \hat{\beta} \) when \( y_j \) is discarded with the asymptotic covariance matrix of \( \hat{\beta} \) (Cook (1977, 1979) in the context of linear regression). That is
\[
b_j = (\hat{\beta}_{j} - \hat{\beta})^T X^T W X (\hat{\beta}_{j} - \hat{\beta})
\]
with \( \hat{\beta}_{j} \) the estimate of \( \hat{\beta} \) with \( y_j \) discarded. The weight matrix \( W \) is evaluated at \( \hat{\beta} \). This is equivalent to comparing the estimates of \( \hat{\beta} \) from (2.10) and (2.1). Following Pregibon (1981), we use for simplicity the one-step scoring estimator \( \hat{\beta}_{j}^{1} \), starting from \( \hat{\beta} \), in place of \( \hat{\beta}_{j} \). We then obtain
\[
b_j^{1} = (\hat{\beta}_{j}^{1} - \hat{\beta})^T X^T W X (\hat{\beta}_{j}^{1} - \hat{\beta})
\]
\[
= \frac{w_j h_j}{\phi_j v_j (1-h_j)^2} (y_j - u_j)^2.
\]

Similarly in the dispersion model we have
which compares the estimates of $\gamma$ from (2.12) and (2.2). We therefore have dual measures of the influence of the $j$th observation, corresponding to its effects on the mean and dispersion estimates respectively.

2.16 Quasi-likelihoods

The class of generalized linear models was derived in §1.7 to have the property that, for any member of the class, the form of the scoring iteration is determined by its mean and variance structure. If the same scoring iteration is used, but based on only second moment rather than full distributional assumptions, the objective function implicitly being maximized is called a quasi-likelihood after Wedderburn (1974). The parameter estimates so obtained are consistent and have similar properties to maximum likelihood estimates, as was shown by McCullagh (1983).

We seek in this section to obtain estimates for both $\hat{\beta}$ and $\gamma$ using only the mean and variance structure assumed in §2.2. For any particular but fixed value of $\gamma$, we can obtain a quasi-maximum likelihood estimate of $\hat{\beta}$. To estimate $\gamma$ in the same way though, with $\hat{\beta}$ fixed, we would have to specify the dispersion variance function $\kappa$. This we prefer not to do, since it involves making at least 4th moment assumptions about the distribution of $\gamma$.

Another possibility would be to use Nelder and Pregibon's extended quasi-likelihood (McCullagh and Nelder, 1983, §11.4.3), which
is in our notation

\[ (2.14) \quad k' = -\frac{1}{2} \sum_i \{ \phi_i^{-1} d'_i + \log(2\pi \phi_i v(y_i)) \}. \]

Here the \( d'_i \) are quasi-deviance components defined by

\[ d'_i = 2\omega_i (Q(y_i, y_i) - Q(y_i, \mu_i)) \]

with \( Q(y, \mu) = \int y - \mu \sqrt{v(\eta)} \, d\eta \). Then \( k' \) depends only on second moment assumptions, and is a close approximation to the true log-likelihood for the distributions commonly used in generalized linear models. It is exactly the likelihood for the normal and inverse Gaussian distributions, but requires some modification for discrete distributions.

Maximizing (2.14) with respect to \( \hat{\beta} \) gives the quasi-maximum likelihood estimate of \( \hat{\beta} \) for fixed \( \hat{\gamma} \). On the other hand, maximizing with respect to \( \hat{\gamma} \) satisfies

\[ \hat{\gamma} = Z^T \left( 2\phi^2 \right)^{-1} (d' - \hat{\phi}) = 0 . \]

This gives maximum likelihood estimation for the normal or inverse Gaussian distributions, but is not generally consistent assuming second moment information only. Indeed it is not generally possible to evaluate \( \mathbb{E}(d') \) without distributional assumptions.

Instead of (2.14), I suggest we assume

\[ (2.15) \quad \hat{k}_\beta = x^T (\phi y)^{-1} \phi \]

\[ (2.16) \quad \hat{k}_\gamma = Z^T \left( 2\phi^2 \right)^{-1} (d' - \hat{\phi}) \]

where now \( d'_i = \omega_i \phi_i \mu_i^{-1} (y_i - \mu_i)^2 \), without explicitly evaluating \( k' \). These equations give maximum likelihood estimation for the normal distribution. For the inverse Gaussian distribution we now have \( d'_i = \omega_i \mu_i^{-3} (y_i - \mu_i)^2 \), whereas maximum likelihood estimation would arise if the \( \phi_1^3 \) was replaced by \( \phi_1^2 \phi_1 \). On the other hand, both \( \hat{k}_\beta \)
and \( \hat{k}_\gamma \) have zero expectation under second moment assumptions, so the estimates obtained by equating them to zero are generally consistent. Differentiating (2.15) and (2.16) shows that \( \hat{\beta} \) and \( \hat{\gamma} \) are quasi-orthogonal and have quasi-information matrices

\[
I'_\beta = X^T WX \quad \text{and} \quad I'_\gamma = Z^T V Z.
\]

If the functions \( g \) and \( h \) are smooth, if \( n^{-1}I'_{\beta} \) and \( n^{-1}I'_{\gamma} \) have positive definite limits, and if \( \gamma \) has finite 3rd to 6th moments, then the asymptotic results of McCullagh (1983) apply to both \( \hat{\beta} \) and \( \hat{\gamma} \). That is, denoting by \( \hat{\beta} \) and \( \hat{\gamma} \) the estimates obtained from \( \hat{k}_\beta = \hat{k}_\gamma = 0 \), and by \( \beta_0 \) and \( \gamma_0 \) the true values,

\[
\mathbb{E}(\hat{\beta} - \beta_0) = O(n^{-1}) \quad \mathbb{E}(\hat{\gamma} - \gamma_0) = O(n^{-1})
\]

\[
n^{-k_2}_{\hat{k}_\beta} \sim N(0, n^{-1}I'_{\beta}) + O_p(n^{-k_2})
\]

\[
n^{-k_2}_{\hat{k}_\gamma} \sim N(0, n^{-1}I'_{\gamma}) + O_p(n^{-k_2})
\]

\[
n^{-k_2}(\hat{\beta} - \beta_0) \sim N(0, n^{-1}I'_{\beta}) + O_p(n^{-k_2})
\]

\[
n^{-k_2}(\hat{\gamma} - \gamma_0) \sim N(0, n^{-1}I'_{\gamma}) + O_p(n^{-k_2})
\]

If the 5th and 6th moments are infinite, the error terms for \( \hat{k}_\gamma \) and \( \hat{\gamma} - \gamma_0 \) are \( O_p(1) \).

An application of the quasi-likelihood methods described above is to fit over-dispersed binomial or Poisson distributions (McCullagh and Nelder, 1983) with varying over-dispersion parameters.
2.17 Software

The computer programs Genstat and GLIM have been designed to handle generalized linear models explicitly. Both programs include many ancillary operations such as plotting, sorting and editing. In this section we give some examples of GLIM commands to implement the suggestions of this chapter. They are not intended as a substitute for a knowledge of the GLIM command language, but to illustrate the ease with which the methods may be implemented.

Suppose we want to fit a normal model with mean and variance structure

\[ \mu = X\beta \quad \log \sigma^2 = Z\gamma. \]

The following are GLIM macros to fit the mean and dispersion submodels respectively. We assume the data is stored in GLIM vector \( Y \), the prior weights in \( W \), the mean covariates in vectors \( X_1 \) to \( X_P \), and the dispersion covariates in vectors \( Z_1 \) to \( Z_Q \).

```
$MACRO MEAN ! MEAN SUBMODEL
$YVAR Y
$ERROR N $LINK I $WEIGHT WS
$CALC WS = W/S ! WEIGHTS = PRIOR WEIGHT/VARIANCE
$FIT X1 + ... + XP
$CALC M = %FV ! STORE FITTED MEANS
$ENDMAC

$MACRO DISPER ! DISPERSION SUBMODEL
$CALC D = %DI ! D = DEVIANCE COMPONENTS
$YVAR D
$ERROR G $LINK L $WEIGHT $SCALE 2
$CALC %FV = S ! START FROM PREVIOUS FITTED VARIANCES
```
The following macro combines the previous two macros into a coupled iteration:

```
$MACRO COUPLE
$RECYCLE
$USE MEAN ! FIT MEAN SUBMODEL
$DISPLAY E ! DISPLAY ESTIMATED BETAS
$USE DISPER ! FIT DISPERSION SUBMODEL
$DISPLAY E ! DISPLAY ESTIMATED GAMMAS
$CALC \%S = \%CU(\%LP) ! N - 2 * LOG-LIKELIHOOD
  : \%C = \%LT(\%R,\%S) ! TEST FOR INCREASE IN LIKELIHOOD
  : \%R = \%S
$ENDMAC
```

Algorithm 3 of §2.11 could then be implemented by a calling sequence such as the following:

```
$CALC S = 1 ! START WITH CONSTANT VARIANCES
$YVAR Y $FIT \%GM $CALC \%R = \%N * \%LOG(\%SC) ! STORE LIKELIHOOD OF NULL MODEL
$WHILE \%C COUPLE ! ITERATE UNTIL CONVERGENCE
```

Differencing the final values of \%R from different models will give overall likelihood ratio tests. To implement Algorithm 2 (alternate iterations) we only need change the $RECYCLE command in macro COUPLE to $RECYCLE 1. Note that this is an example for which the mean model converges in one iteration, so that Algorithms 2 and 3 each require fewer iterations than the method of scoring.

As pointed out by Pregibon (1982), score tests are also easily
calculated in GLIM using the \$X2$ system scalar. Suppose for example that we have fitted a model using the macro \texttt{COUPLE} and wish to test if another dispersion covariate, \texttt{ZR} say, should be included. Then the sequence

\begin{verbatim}
$CALC \%C = \%X2$
$RECYCLE 1$
$FIT + ZR$
$CALC \%C - \%X2$
\end{verbatim}

will output the score test of $\gamma_r = 0$. 
3.1 Introduction

In this chapter we consider several ways to go outside the extended generalized linear model framework of chapter 2. For simplicity we restrict ourselves to the normal distribution with unit prior weights. We will suppose that

\[ y_i \sim N(\mu_i, \sigma_i^2) \quad i = 1, \ldots, n \]

and will write

\[ d_i = (y_i - \mu_i)^2. \]

Our intention is merely to state some possibilities and make some suggestions. Section 3.2 makes the observation that any nonlinear regression model may be applied in each of the submodels. Section 3.3 suggests the use of smoothing splines as a nonparametric method. Section 3.4 considers a model for which the variance is functionally related to the mean. Section 3.5 addresses the question of whether to model a changing variance or whether to attempt to remove any trend by a transformation of the data.

3.2 Nonlinear regression

The dependence of \( \mu \) on parameters \( \beta \), and \( \sigma^2 \) on parameters \( \gamma \), may be quite general without disturbing the orthogonality of \( \beta \) and \( \gamma \). Write \( \hat{\sigma}^{-1} = (\hat{\sigma}^2) \) and \( \hat{\sigma}^{-1} = (2\hat{\sigma}^4) \), and let \( \dot{\mu} \) and \( \dot{\sigma}^2 \) represent differentiation with respect to \( \beta \) and \( \gamma \) respectively.
The scoring iterations in the mean and variance submodels are

\[ \hat{\beta}^{k+1} = \hat{\beta}^k + (\hat{\mu}^T \hat{\sigma}^2)^{-1} \hat{\mu}^T (y - \mu) \]

and

\[ \hat{\gamma}^{k+1} = \hat{\gamma}^k + (\sigma^2 \hat{V}_0)^{-1} \sigma^2 \hat{V} (d - \sigma^2) \]

respectively. Comparing with chapter 2, each of the submodels now pose an arbitrary nonlinear regression problem, rather than the limited nonlinearity of a generalized linear model. Coupled and alternate iterations still provide methods for organizing the likelihood calculations.

3.3 Nonparametric Submodels

Return for a moment to constant variance regression, and consider the model

\[ \mu_i = g(t_i), \quad \sigma_i^2 = \sigma^2, \quad t_i \in [0,1] \]

where \( g \) is an unknown function. When fitting such a model in practice, it may well be the case that we are unwilling to assume a specific parametric form for \( g \). But suppose that \( g \) is smooth in the sense that it has, for some \( p, p-1 \) continuous derivatives and

\[ (3.1) \quad \int_0^1 (g^{(p)}(t))^2 dt \]

is small (\( p = 2 \) corresponds to visual smoothness). Then an appropriate objective function for \( g \) is a weighted sum of the sum of squares and the "roughness" (3.1),

\[ (3.2) \quad \frac{1}{n} \sum_i (y_i - g(t_i))^2 + \lambda \int_0^1 (g^{(p)}(t))^2 dt. \]

Minimizing (3.2) with respect to \( g \) leads to fitting a polynomial spline of order \( 2p-1 \) (Schoenberg, 1964). Reinsch (1967) gives a
computational algorithm for \( p = 2 \), which basically involves solving a banded linear \( n \times n \) system for the quadratic parameters. Wahba (1978, 1983) shows that this is a Bayesian posterior mean, given a prior which is diffuse on the coefficients of the polynomials of degree \(< 2p-1\). She persuasively argues that spline smoothing therefore provides a natural generalization of least squares regression.

The relative weight parameter \( \lambda \) may be estimated by a cross-validation procedure such as that of Craven and Wahba (1979). Spline smoothing using this estimate then provides a consistent nonparametric estimation procedure for the means \( \mu_i \).

Our interest in spline smoothing is that it may be applied to the variance submodel. Suppose that

\[
\sigma_i^2 = h(t_i)
\]

where \( h \) is an unknown function. It will often be the case that we wish to hedge against the variance changing in an unknown way, even if we are prepared to assume a parametric function for the means. If \( h \) is smooth in the sense used above, it is appropriate to minimize

\[
\frac{1}{n} \sum_i (d_i/h(t_i)) + \log h(t_i)) + \lambda \int_0^1 (h^{(p)}(t))^2 dt.
\]

Comparing (3.3) with (3.2), the kernel of the gamma distribution has displaced the sum of squares. The minimum still occurs at a polynomial spline, the calculation of which now involves solving a nonlinear banded \( n \times n \) system. Cross-validation is still available to estimate \( \lambda \). I conjecture that the estimated spline may still be interpreted as a posterior mean.
3.4 Mean and Variance Models Overlap

Suppose that the variance is functionally related to the mean. Consider the model

\[ g(\mu) = x_1 \delta + x_2 \delta \]
\[ h(\sigma^2) = z_1 \delta + z_2 \gamma \]

for which the mean and variance submodels share parameters. Two approaches to this model are suggested. Firstly we may make the mean and variance submodels distinct by putting \( \delta \) to be \( \delta_1 \) in the former and to be \( \delta_2 \) in the latter, \( \delta_1 \) and \( \delta_2 \) not necessarily being equal. Coupled or alternate iterations may be used to estimate \( \delta_1, \gamma, \delta_2 \). Whereupon solving

\[ (x_2^T w x_2 + z_1^T v z_1) \delta = x_2^T w x_2 \delta_1 + z_1^T v z_1 \delta_2 \]

with \( W = \langle g^2 / \sigma^2 \rangle \) and \( V = \langle h^2 / \sigma^4 \rangle \), yields efficient estimates of all parameters.

Alternatively, full maximum likelihood estimates may be calculated. The score vectors are

\[ k_\beta = x_1^T w (y - \mu) \]
\[ k_\delta = x_2^T w (y - \mu) + z_1^T v (d - \sigma^2) \]
\[ k_\gamma = z_2^T v (d - \sigma^2) \]

and the information matrices are

\[ I_\beta = x_1^T w x_1 \]
\[ I_\delta = x_2^T w x_2 + z_1^T v z_1 \]
\[ I_\gamma = z_2^T v z_2 \].
The scoring algorithm for all parameters simultaneously may be implemented by constructing a generalized linear model with data \( \begin{pmatrix} y \\ d \end{pmatrix} \), covariates \( \begin{pmatrix} x_1 & x_2 & 0 \\ 0 & z_1 & z_2 \end{pmatrix} \), link derivatives \( \begin{pmatrix} g \\ h \end{pmatrix} \), variance function \( \sigma^4 \) and dispersion vector \( \begin{pmatrix} \gamma^2 \\ \beta^2 \end{pmatrix} \).

3.5 Transform or Model the Variance?

Suppose that we are given a data set for which the variance is not constant. The same issue may well arise, whether to treat the data as normal and model the changing variance, or whether to transform and treat the data as normal and homogeneous on the transformed scale. As an example, we consider the alternative hypotheses

\[ H_f : \log y_i \sim N(\xi_i, \delta^2_i) \]
\[ H_g : y_i \sim N(\mu_i, \sigma^2_i) \]

Here the \( \mu_i, \sigma^2_i, \xi_i \) and \( \delta^2_i \) are understood to be parametrized in terms of covariates and explanatory parameters, and the \( \delta^2_i \) will usually be equal. The difficulty of discriminating between such models is that they are separate, in the sense that neither is a special case of the other. The approach of Cox (1961, 1962) and Atkinson (1970) to such problems is to construct a super-model with likelihood proportional to

\[ L_{\lambda} = L_f^\lambda L_g^{1-\lambda} \quad 0 \leq \lambda \leq 1 \]

where \( L_f \) and \( L_g \) are the likelihoods corresponding to \( H_f \) and \( H_g \) respectively. Testing \( H_f \) against \( H_g \) is then equivalent to testing \( \lambda = 0 \) against \( \lambda > 0 \).

Let \( \alpha \) represent \( \xi \) and \( \delta^2 \), and \( \beta \) represent \( \mu \) and \( \sigma^2 \).
Let \( l_f \) and \( l_g \) be the log-likelihoods, let \( l_{fg} = l_f - l_g \),

\[
m(\alpha, \beta) = E_f(l_{fg})
\]

and

\[
v(\alpha, \beta) = D_f(l_{fg}) - E_f(l_{fg}^T)D_f(l_f)E_f(l_{fg})
\]

where \( E_f \) and \( D_f \) represent mean and dispersion under \( H_f \). Note that \( v(\alpha, \beta) \) is approximately the conditional variance of \( l_{fg} \) given \( l_f \). Straightforward calculation gives the score test of \( \lambda = 0 \) as

\[
v^{-\frac{1}{2}}(l_{fg} - m).
\]

This test is nonstandard since \( \lambda = 0 \) is on the boundary of the set of permissible values. Moreover, it is not at all clear what it seems to estimate \( \beta \) under \( H_f \). Cox (1962) uses the statistic

\[
v(\hat{\alpha}, \hat{\beta}, \hat{\alpha}) - \nu(\alpha, \beta) - m(\hat{\alpha}, \hat{\beta}) \]

where \( \hat{\alpha} \) and \( \hat{\beta} \) are maximum likelihood estimates under \( H_f \) and \( H_g \) respectively, and \( \hat{\beta} \) is the probability limit of \( \hat{\beta} \) under \( H_f \) evaluated at \( \hat{\alpha} \). Atkinson (1970) uses

\[
v(\hat{\alpha}, \hat{\beta}, \hat{\alpha}) - \nu(\hat{\alpha}, \hat{\beta}) - m(\hat{\alpha}, \hat{\beta}) \]

Another possibility is

\[
v(\hat{\alpha}, \hat{\beta}) - \nu(\hat{\alpha}, \hat{\beta}) - m(\hat{\alpha}, \hat{\beta}) \]

For our example we have, after some calculation,

\[
S_C = l_{fg}(\hat{\alpha}, \hat{\beta}) - m(\hat{\alpha}, \hat{\beta})
\]

\[
= l_{fg}(\alpha, \beta) + \log \sigma^2_A - \log \sigma^2_{A1}
\]

\[
S_A = l_{fg}(\alpha, \beta) - m(\hat{\alpha}, \hat{\beta})
\]

\[
= l_{fg}(\alpha, \beta) - m(\hat{\alpha}, \hat{\beta}) \]

and
\[ S_S = \frac{n}{2} - k \sum_{i=1}^{n} \frac{1}{\sigma_i} \left( \frac{\sigma_i^2}{\mu_i} + (\mu_i - \hat{\mu}_i)^2 \right) \]

with

\[ \mu_{\alpha} = e^{\xi + \delta \sigma^2}, \quad \sigma_{\alpha}^2 = e^{2\xi + \delta^2 (\sigma^2 - 1)} . \]

I have calculated \( v(\alpha, \beta) \), but do not give it here because it is extremely tedious. All three statistics are consistent. In fact, the contributions to \( S_C, S_A \) and \( S_S \) corresponding to a datum \( y \), have probability limits under \( H_g \) which can be calculated to be

\[ - \frac{3}{2} \frac{\sigma^2}{\mu^2} - \frac{59}{12} \frac{\sigma^4}{\mu^4} , \]

\[ - \frac{3}{2} \frac{\sigma^2}{\mu^2} - \frac{8}{3} \frac{\sigma^4}{\mu^4} , \]

and

\[ - \frac{3}{2} \frac{\sigma^2}{\mu^2} - \frac{43}{6} \frac{\sigma^4}{\mu^4} \]

respectively. Here we have assumed that \( \frac{\sigma^2}{\mu^2} \) is small enough that the probability of a negative observation is negligible, and have neglected cubic and higher terms in it. The corresponding test of \( H_g \) against \( H_f \) may be constructed from the score test of \( \lambda = 1 \), but is not given here.

Other approaches to testing separate hypotheses are developed by Borth (1975), Fisher and McAleer (1981), Epps et al. (1982) and Atkinson (1982), but seem no easier to calculate than the test given above. The test given here must be considered cumbersome to calculate and difficult to interpret. Furthermore its convergence to a standard normal statistic is very slow (see Jackson, 1968, and Pereira, 1977). I would recommend that in practice the discrimination between a normal and a log-normal distribution be made on the basis of whether the errors have a skew distribution and whether the data is structurally positive.
CHAPTER 4
MULTIVARIATE AND AUTOREGRESSIVE MODELS

4.1 Introduction

Suppose that \( y \in \mathbb{R}^n \) has a multinormal distribution with mean \( \mu \) and covariance matrix \( \Sigma \). Let

\[
\Sigma = LDL^T
\]

with \( D \) diagonal and \( L \) unit lower triangular. Also let \( \lambda_{i1}, \ldots, \lambda_{i,i-1} \) be the partial regression coefficients of \( y_i \) on \( y_1, \ldots, y_{i-1} \), and \( \sigma_i^2 \) be the corresponding conditional variances. So

\[
E (y_i - \mu_i | y_1, \ldots, y_{i-1}) = \sum_{j=1}^{i-1} \lambda_{ij} (y_j - \mu_j)
\]

and

\[
D (y_i | y_1, \ldots, y_{i-1}) = \sigma_i^2.
\]

The factorization of the covariance matrix gives

\[
L^{-1}(y - \mu) \sim N(0, D)
\]

so it must be that \( L_{ij} = -\lambda_{ij} \), where \( L_{ij} \) is an element of \( L^{-1} \), and \( D_{ii} = \sigma_i^2 \). This identity is used in this chapter to fit varying parameter autoregressive and multinormal models. Our development is very brief and will serve only to point out the possibilities.

4.2 Autoregression

4.2.1 A varying parameter model

Let \( y \) be multinormal as in the last section. It is our
intention to allow the means, partial regression coefficients and conditioned variances to vary with the index $i$. We do this by assuming the following parametric model,

$$
\mu = X\alpha \\
\log \sigma^2 = Z\gamma \\
I - L^{-1} = \sum_{j=1}^{q} \beta_j B_j
$$

with $\alpha \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}^q$. Here $X$, $Z$ and the $B_j$ are known covariate matrices.

We can now go ahead and calculate the necessary conditions for a maximum of the likelihood. The log-likelihood is

$$
\ell = -\frac{1}{2} \log |\Sigma| - \frac{1}{2}(y - \mu)^T \Sigma^{-1} (y - \mu) \\
= -\frac{1}{2} \log |D| - \frac{1}{2}(y - \mu)^T L^{-1} D^{-1} L^{-1} (y - \mu).
$$

It has derivatives

$$
\dot{\ell}_\alpha = X^T L^{-1} (y - X\alpha) \\
\dot{\ell}_\beta_j = (y - \mu)^T B_j D^{-1} L^{-1} (y - \mu) \\
\dot{\ell}_\gamma = \frac{1}{2} 2^T (D^{-1} d - l)
$$

with $d = (L^{-1} (y - \mu))^2$. We can express $\dot{\ell}_\beta$ as

$$
\dot{\ell}_\beta = y^T D^{-1} (\gamma - Y\beta)
$$

where $\gamma = y - \mu$ and the $j$th column of $Y$ is $B_j (y - \mu)$. We see that the parameter sets $\alpha$, $\beta$ and $\gamma$ are mutually orthogonal, since the second derivative are all linear in $y - \mu$, and hence have expectation zero. Also, solving $\dot{\ell}_\alpha = 0$ is equivalent to regressing $y$ on $X$ with weight matrix $\Sigma$, solving $\dot{\ell}_\beta = 0$ is equivalent to regressing $y$ on $Y$ with weight matrix $D$, and solving $\dot{\ell}_\gamma = 0$ is
equivalent to regressing \( \tilde{d} \) on \( Z \) with gamma errors and log-link.

So we propose the following coupled iteration sequence, which has the overall maximum likelihood estimates of \( \tilde{\alpha}, \tilde{\beta}, \) and \( \tilde{\gamma} \) as a stationary point.

repeat

\[
\begin{align*}
\text{regress } \tilde{y} \text{ on } X \text{ with weight matrix } \Sigma \text{ to update the estimate of } \tilde{\alpha}. \\
\text{regress } y \text{ on } Y \text{ with weight matrix } D \text{ to update the estimate of } \tilde{\beta}. \\
\text{fit a generalized linear model to } \tilde{d} \text{ with gamma errors, log-link, covariate matrix } Z, \text{ and scale parameter } \tilde{\gamma}, \\
\text{to update the estimate of } \tilde{\gamma}. \\
\end{align*}
\]

until convergence.

4.2.2 An example

Consider the autoregressive model

\[
y_i - \mu_i = \sum_{j=1}^{s} \lambda_{ij}(y_{i-j} - \mu_{i-j}) + \epsilon_i
\]

with \( \epsilon_i \sim N(0, \sigma_i^2) \) and \( y_i = \mu_i = 0 \) for \( i \leq 0 \). A simple linear trend model for the autoregressive parameters is

\[
\lambda_{ij} = \beta_{ij} + \beta_{2j} i
\]

The component of the above algorithm which updates the estimate of \( \tilde{\beta}, \) takes the form then of regressing \( \tilde{y} = (y - \tilde{y}) \) on
4.3 Multinormal Models

Let \( y_i \sim N(\mu_i, \Sigma_i) \) for \( i = 1, \ldots, m \). Let \( (y_{i1}, \ldots, y_{im})^T \) have covariance matrix \( \Sigma \). It is our intention to allow \( \Sigma \) to vary with \( i \). To do this we have the \( LDL^T \) decomposition of \( \Sigma \) to reduce the problem to a sequence of \( m \) univariate regressions.

Consider the conditional regressions

\[
\mathbb{E}(y_1) = \mu_1 \\
\mathbb{E}(y_2 | y_1) = \mu_2 + \lambda_{21}(y_1 - \mu_1) \\
\mathbb{E}(y_3 | y_1, y_2) = \mu_3 + \lambda_{31}(y_1 - \mu_1) + \lambda_{32}(y_2 - \mu_1)
\]

and so on. If linear predictors are introduced for the \( \mu_j \) and the \( \lambda_{jk} \), then the methods of chapter 2 may be applied to each of the \( m \) conditional regressions.
PART II
SEPARABLE REGRESSION AND PRONY'S PARAMETRIZATION

PREAMBLE

The material of part II is motivated by the problem of fitting sums of exponentials

\[ \mu(t) = \sum_j \alpha_j e^{-\beta_j t} \]

to equispaced data by least squares. This problem is well known to be numerically difficult (Varah, 1985). In particular, general purpose algorithms, such as Gauss-Newton and its variants, often have great difficulty in converging. Yet sums of exponentials arise quite frequently, usually as the general solutions of homogeneous constant coefficient differential equations

\[ \sum_j \xi_j \mu^{(j-1)} = 0. \]

For \( \mu \) evaluated at equispaced times, there are corresponding constant coefficient difference and recurrence equations.

We develop an algorithm, originally due to Osborne (1975), that explicitly uses the fact that \( \mu \) satisfies a difference equation. Doing so involves two steps. We separate out the linear parameters, as described in §1.13. And we reparametrize from the \( \beta_j \) to the coefficients \( \gamma_j \) of the difference equation, as was first done by Prony in 1795. We then find that the derivative of the reduced objective function can be written as \( B(\gamma) \gamma \), where \( B \) is a symmetric
matrix function of $\gamma$. Osborne's algorithm, which we called modified Prony, arises by treating

$$B(\gamma)\gamma = 0$$

as a nonlinear eigenproblem. The iterative process is defined by letting $\gamma^{k+1}$ be the eigenvector of $B(\gamma^k)$ with eigenvalue nearest zero.

In fact our development is more general, as it applies to any function satisfying a differential equation with coefficients linear and homogeneous in the parameters. The more general class includes rational functions. The main thread of our development is directed towards proving the asymptotic stability of the modified Prony algorithm. Chapter 5 introduces the problem, specifies the assumptions under which asymptotic results will be proved, and arrives at the nonlinear eigenproblem formulation. Chapter 6 describes the resulting algorithm, including some computational details. Chapter 7 proves the stability result, that the convergence factor of the modified Prony algorithm has almost sure limit zero. Some resort is made in this proof to the special structure of the exponential and rational fitting examples. Chapter 8 compares the convergence of the modified Prony and Gauss-Newton algorithms by way of simulations.
5.1 Introduction

In this chapter we set ourselves a nonlinear least squares problem arising from sampling a continuous process at equispaced times. Suppose the continuous time process \( \{y(t) : t \geq 0\} \) has expectation \( \{\mu(t;\theta_0) : t \geq 0\} \), where, for each \( \theta \), \( \mu(t;\theta) \) is a continuous function of \( t \). Suppose the process is sampled at equispaced times \( t_i \) to obtain observations \( y_i = y(t_i) \) with means \( \mu_i = \mu(t_i;\theta_0) \). Without essential loss of generality we assume that, for any sample size \( n \), the \( t_i \) are spaced over the unit interval with \( t_i = \frac{i}{n} \). Suppose further that the errors \( y_i - \mu_i \) may be considered independent and identically distributed. In the absence of further distributional assumptions, it is appropriate to estimate \( \theta \) by minimizing the sum of squares

\[
\phi(\theta) = (y - \mu)^T(y - \mu).
\]

In the next section, regularity conditions similar to those assumed in §1.15 are shown to be sufficient to prove consistency of the least squares estimates and asymptotic stability of the Gauss-Newton algorithm. In section 5.3 some properties of circulant matrices are summarized. Sections 5.4 and 5.5 reduce the least squares problem to a nonlinear eigenproblem, assuming a difference equation structure for \( \mu \). Sections 5.6 and 5.7 relate the difference equation formulation to recurrence and differential equations.
Sections 5.8 and 5.9 introduce the examples of exponential and rational fitting.

5.2 The Asymptotic Sequence

The asymptotic assumptions and results of §1.15 cannot be applied directly to our least squares problem because $t_1, \ldots, t_n$ do not appear as the leading $n$ terms of an infinite sequence. In fact they cannot do so if we require them to be equispaced in a fixed interval. The purpose of this section is to deal with this point. To that end, details of the asymptotic sequence are introduced. Having been used to establish two theorems which will be used later, these details may be forgotten.

In this section only, we will make the dependence of $t_i$ on $n$ explicit by writing $t_{n,i}$. Throughout part II we will assume that, for any $n$, the $t_{n,i}$ are in ascending order with $t_{n,i} = i/n$. This might be called time ordering. In order to construct an asymptotic sequence, we introduce the series $\{s(i) : i = 1, 2, \ldots\}$ equal to

$$1, \frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \ldots$$

That is $s(i) = (2k+1)/2^m$ with $i = 2^{m-1} + k + 1$, $m \geq 0$, $k \geq 0$, $k < 2^{m-1}$. For any $n$, let $t_{n(1)}, \ldots, t_{n(n)}$ be equal to $s(1), \ldots, s(n)$ slightly perturbed to make them equally spaced. For example, the $t_{n(i)}$'s are $1, \frac{2}{3}, \frac{1}{3}$ for $n = 3$ and $1, \frac{1}{2}, \frac{1}{4}, \frac{3}{4}$ for $n = 4$. Then $t_{n(1)}, \ldots, t_{n(n)}$ is simply a reordering of $t_{n,1}, \ldots, t_{n,n}$. We call this asymptotic sequence ordering; it is to be thought of as indicating the way in which the sampling grid is refined as more observations are sought. The $t_{n(i)}$ have the $s(i)$ as limiting values, $t_{n(i)} = s(i) + 0(n)$ uniformly as $n \to \infty$. In
Let $y_{n(i)}$ and $\mu_{n(i)}$ represent the observations and means in asymptotic sequence order. We make the basic assumption that the errors $\varepsilon_{n(i)} = y_{n(i)} - \mu_{n(i)}$ do not depend on $n$. Our strategy will be to apply the results of Jennrich (1969) (see §1.15) to the least squares problem of minimizing

$$
\frac{1}{n} \sum_{i=1}^{n} (\varepsilon_{n(i)} - \mu_{n(i)}(\theta) - \mu_{n(i)}(\theta))^2
$$

and extend them to minimizing

$$
\frac{1}{n} \sum_{i=1}^{n} (\varepsilon_{n(i)} - \mu_{n(i)}(\theta) - \mu_{n(i)}(\theta))^2
$$

by continuity of $\mu$. In this context, assumptions (a) to (d) of §1.15 may be written as follows. Note that limiting sums of products, of $\mu$ and its derivatives, here become integrals over the unit interval.

(a) The errors $\varepsilon_{n(i)}$ are independent and identically distributed with mean 0 and finite variance $\sigma^2$. The parameter space $\theta$ is a compact subset of $\mathbb{R}^k$, and contains $\theta_0$ as an interior point.

(b) The function

$$
Q(\theta) = \int_0^1 (\mu(t; \theta_0) - \mu(t; \theta))^2 dt
$$

has a unique minimum in $\theta$ at $\theta_0$.

(c) The function $\mu$ is twice continuously differentiable with respect jointly to $\theta$. All these derivatives are continuous in $t$.

(d) The information matrix $H$, defined by
\[
H_{ij} = \int_0^1 \frac{\partial^2 u}{\partial \theta_i \partial \theta_j} (t; \theta) \ dt
\]
is positive definite.

We are now in a position to prove the following results.

**Theorem 5.1 (Law of large numbers).** If, for every \( \theta \in \Theta \), \( f(t; \theta) \) is a continuous function on the unit interval, then

\[
\frac{1}{n} \sum_{i=1}^{n} f(t_i; \theta) \to \text{a.s.} \quad 0
\]
uniformly over \( \Theta \).

**Proof.** For fixed \( \theta \), the result follows from Theorem 4.1.3 of Stout (1974)* or Corollary 1 of Stout (1968)*. The proof that the result holds uniformly follows closely that of Theorem 4 of Jennrich (1969).

**Theorem 5.2.** The least squares estimator \( \hat{\theta} \) is strongly consistent.

**Proof.** Follows closely that of Theorem 6 of Jennrich (1969).
Armed with the law of large numbers and consistency, the asymptotic stability of the Gauss-Newton algorithm may be proved exactly as in §1.15.

Remark. An alternative asymptotic sequence would have arisen had we assumed a triangular array of independent errors \( (\epsilon_{n,i} : 1 \leq i \leq n) \). We would then have needed to prove that

\[
\frac{1}{n} \sum_{i=1}^{n} f(t_{n,i}; \theta) \epsilon_{n,i} \overset{a.s.}{\to} 0
\]

As remarked by Stout (1968)*, the generalization to this case is trivial because the same proofs remain valid without modification.


5.3 Circulant Matrices

It is convenient, here at the outset, to summarize briefly the properties of circulant matrices, which will yield considerable service in the sequel. For more details see Davis (1979). An \( n \times n \) matrix is *circulant* if it has the form
which may also be written \( C = \text{circ}(c_1, c_2, \ldots, c_n) \). Write

\[
P_c(z) = \sum_{i=1}^{n} c_i z^{i-1}
\]

and let

\[
\omega = e^{\frac{2\pi}{n}\sqrt{-1}}
\]

the fundamental \( n \)th root of unity. Define the \( n \times n \) Fourier matrix \( F \) by

\[
F_{ij} = \omega^{(i-1)(j-1)}
\]

which is a unitary circulant matrix. The basic reason for the usefulness of circulant matrices, is that they are all diagonalized by the Fourier matrix of the same size, with (complex) eigenvalues

\[
\lambda_i = P_c(\omega^{i-1}) \quad i = 1, \ldots, n
\]

That is, we can write

\[
C = F^* \Lambda F
\]

with \( \Lambda = \{\lambda_i\} \). Following the evocative language of signal processing, we call \( F_z \) the discrete Fourier transform of \( z \), \( F_z^* \) the inverse discrete Fourier transform, and \( P_c(\cdot) \) the transfer function of the matrix \( C \).

Of particular interest is the circulant forward shift operator \( \Pi = \text{circ}(0,1,0,\ldots,0) \). An alternative characterization of circulant matrices is as linear combinations of powers of \( \Pi \).
5.4 Difference Equations

We will assume that \( y(t) \) satisfies a difference equation

\[
(5.3) \quad \sum_{k=1}^{p+1} d_k(t;\gamma) \Delta^{k-1} y(t) = 0
\]

where \( \Delta \) represents the forward difference operator

\[
\Delta f(t) = n \left( f \left( t + \frac{1}{n} \right) - f(t) \right).
\]

and the coefficients \( d_k \) are continuous in \( t \) and linear and homogeneous in \( \gamma \in \mathbb{R}^{q+1} \). Generally \( \gamma \) and \( d \) will depend on \( n \), but we will usually not make this explicit in our notation. Since the scales of \( \gamma \) and \( d \) are disposable, we will adjoin the condition \( \gamma^T \gamma = 1 \) or \( \gamma_{q+1} = 1 \) as convenient, and similarly for \( d \). By assumption, the \( d_k \) may be expanded to

\[
d_k(t;\gamma) = \sum_{j=1}^{q+1} \gamma_j d_{kj}(t)
\]

where the \( d_{kj} \) are continuous functions, so (5.3) may be rewritten as

\[
(5.4) \quad \sum_{j=1}^{q+1} \sum_{k=1}^{p+1} \gamma_j \left( \sum_{k=1}^{p+1} d_{kj}(t) \Delta^{k-1} \right) y(t) = 0.
\]

In matrix terms, (5.4) becomes the following. Let \( \Delta \) now represent the \( n \times n \) circulant difference matrix \( \Delta = n(\Pi - I) \), and \( P \) be the \( (n-p) \times n \) matrix \( (I_{n-p} \ 0) \). Then \( \mu \) satisfies

\[
(5.5) \quad P \sum_{j=1}^{p+1} \gamma_j \left( \sum_{k=1}^{q+1} \left( d_{kj}(\gamma) \right) \Delta^{k-1} \right) \mu = 0.
\]

Write also

\[
C_j = \sum_{k=1}^{p+1} \left( d_{kj}(\gamma) \right) \Delta^{k-1}
\]

and
Let $X_j = C_j^T$, the leading $n-p$ columns of $C_j$, and let

$$X = \sum_{j=1}^{q+1} \gamma_j X_j.$$ 

Then (5.5) may be re-expressed as

$$PC(\gamma)\mu = 0$$

or

$$X(\gamma)\mu = 0.$$ 

Note that $X$ is a $n \times (n-p)$ matrix with $p+1$ bands. It has full rank if $d_{p+1} \neq 0$. We call $\gamma$ the Prony parameters.

5.5 The Normal Equations as a Nonlinear Eigenproblem

Let

$$\beta = QC(\gamma)\mu,$$

where $Q$ is the $p \times n$ matrix $(0 I_p)$. Then $\mu$ may be parametrized in terms of $\gamma$ and $\beta$ as

$$\mu = C(\gamma)^{-T}Q^T\beta.$$ 

Write $A = C^{-T}Q^T$; the $p$ columns of $A$ correspond to particular solutions of the difference equation (5.3). In this way we have displayed $\mu$ as a separable regression, with $\beta$ and $\gamma$ as the linear and nonlinear parameters respectively. As described in §1.13, the least squares problem may be solved by minimizing

$$\psi(\gamma) = \beta^T(I - P_A)\gamma$$

with respect to $\gamma$, where $P_A$ is the projection onto $R(A)$, and recovering the least squares estimate of $\beta$ from
\[ \hat{\alpha}(\gamma) = (A^T A)^{-1} A^T y. \]

Note that \( X^T A = 0 \), so that \( X \) and \( A \) span orthogonal spaces.

Therefore (5.7) can be written more conveniently as

\[ \psi(\gamma) = y^T P \gamma. \]

Since \( P = X(X^T X)^{-1} X^T \), \( \psi \) has partial derivatives

\[ \psi_i = 2y^T X_i (X^T X)^{-1} X_j \gamma - 2y^T X (X^T X)^{-1} X_i X_j X^T \gamma \]

\[ = 2 \sum_{j=1}^{q+1} \gamma_j B_{ij} \]

with

\[ B_{ij} = y^T X_i (X^T X)^{-1} X_j \gamma - y^T X (X^T X)^{-1} X_i X_j (X^T X)^{-1} X^T \gamma. \]

Therefore, gathering the \( B_{ij} \) into a symmetric matrix function of \( \gamma \),

\[ \dot{\psi} = 2B(\gamma) \gamma. \]

A necessary condition for a minimum of the least squares problem is therefore

\[ B(\gamma) \gamma = 0. \]

We call \( B \) the Prony matrix. The interpretation and solution of (5.9) as a nonlinear eigenproblem is the subject of chapter 6.

The class of Prony parametrizations, for which the normal equations can be expressed in the form (5.9), are related by linear transformations. Apart from a scale factor, the matrices \( C \) and \( X \) are invariant under reparametrization, but \( C_j = \frac{\partial C}{\partial \gamma_j} \) and \( X_j = \frac{\partial X}{\partial \gamma_j} \) are not. If two parametrizations \( \gamma \) and \( \delta \) are related by \( U \delta = \gamma \), where \( U \) is a nonsingular matrix, then their Prony matrices are related by...
5.6 Recurrence Equations

It is convenient to reformulate (5.3) as a recurrence equation

\[(5.11) \sum_{k=1}^{p+1} c_k(t; \gamma) \Pi_{j=1}^{k-1} \mu(t) = 0\]

with \(\Pi f(t) = f(t + \frac{1}{n})\), because the values of the coefficients give the elements of X. Polynomials in \(A\) satisfy the identity

\[
\sum_{k=1}^{p+1} d_k A^{k-1} = \sum_{k=1}^{p+1} d_k n^{k-1} (\Pi - I)^{k-1}
\]

So, at any time \(t\), the \(c_k\) are given by

\[(5.12) \quad c_j = \sum_{k=j}^{p+1} (-1)^{k-j} n^{k-1} d_k\]

That is, \(c = Ud\) with

\[
U = \begin{pmatrix}
1 & -1 & 1 & \ldots & (-1)^{p-1} & 1 & \ldots & 1 & 1 \\
1 & -2 & 1 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
1 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
1 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
1 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
1 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
1 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
1 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
1 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\end{pmatrix}
\]

\[
D = \begin{pmatrix}
p \\
p \ldots \\
p \ldots \\
p \ldots \\
p \ldots \\
p \ldots \\
p \ldots \\
p \ldots \\
p \ldots \\
\end{pmatrix}
\]

Let \(X\) be given by
(Remember that $X$ is a $n \times (n-p)$ window of $C$.) Its elements are given by

$$c_{k,i}(\gamma) = c_k(t_i;\gamma).$$

It is interesting to note that $U^T$ can be written as

$$U^T = (P + Q_1 A^T) \ldots (P + Q_p A^T)$$

with

$$P_j = \begin{pmatrix} I_j & 0 \\ 0 & 0 \end{pmatrix}, \quad Q_j = I - P_j$$

as it can be calculated by repeated windowed differencing.

5.7 Differential Equations

The derivation of (5.9) depended on $\mu(t)$ satisfying exactly a difference equation. Yet $\mu(t)$ will often come to us as the general solution of a differential equation, say

$$\sum_{k=1}^{p+1} b_k(t_i;\xi) D^{k-1} \mu(t) = 0$$

where $D$ is the differential operator. In fact we will assume that there is such a differential equation, with the $b_k$ continuous in $t$ and linear and homogeneous in $\xi \in \mathbb{R}^{q+1}$, and that the difference
equation (5.3) is a discrete analogue with $y \rightarrow \xi$ and $d \rightarrow b$ as $n \rightarrow \infty$. The import here of assumption (d) of §5.2 is that the matrix with elements

$$
\int_0^1 \left. \frac{\partial u}{\partial x_i} \right|_{t; x_0} \frac{\partial u}{\partial x_j} \left. \right|_{t; x_0} dt
$$

is positive semi-definite, and that the zero eigenvalue, which corresponds to the scale of $\xi$, is isolated with multiplicity one.

We can conclude that asymptotically

$$\frac{1}{n} \mu'_Y (y_0) \mu'_Y (y_0)$$

is positive semi-definite with one isolated zero.

Two important classes of linear homogeneous differential equations for which there are corresponding exact difference equations are dealt with in the next two sections. Even when no exact difference equation exists, it may well be possible to use

$$
P+1 \sum_{k=1}^{p} b_k (t; \xi) \Delta_s^{k-1} \mu (t) = O(n^{-2})
$$

(5.16)

where $\Delta_s = \frac{1}{2} (\Delta + \Delta^T)$ is the symmetric difference operator.

Approximate estimates could be obtained by equating (5.16) to zero, although the consequences of this approach will not be pursued here.

5.8 Exponential Fitting

Suppose that $\mu (t)$ satisfies a differential equation with constant coefficients

$$
P+1 \sum_{k=1}^{p} \xi_k \Delta_s^{k-1} \mu (t) = 0.
$$

(5.17)
Then (5.17) has the form (5.15) with $p = q$ and $b_k(t; z) = \xi_k$. Let the polynomial

$$p(z) = \sum_{k=1}^{p+1} \xi_k z^{k-1}$$

have distinct roots $-\beta_j$ with multiplicities $m_j$, $j = 1, \ldots, s$. Then (5.17) may be rewritten as

$$\prod_{j=1}^{s} (D + \beta_j I)^{m_j} u(t) = 0$$

and the general solution for $u$ is

$$u(t) = \sum_{j=1}^{s} \sum_{k=1}^{m_j} a_{jk} t^{k-1} e^{-\beta_j t}.$$ 

If a pair of complex roots appear, then use the damped trigonometric functions

$$e^{-\text{re}(\beta_j) t} \sin(\text{im}(\beta_j) t), \quad e^{-\text{re}(\beta_j) t} \cos(\text{im}(\beta_j) t).$$

Now it is easy to verify that the operator $(\Delta + \zeta_j I)^m$, with $\zeta_j = n(1 - e^{-\beta_j / n})$, annihilates the term $t^{n-1} e^{-\beta_j t}$ as does $(D + \beta_j I)^m$. Therefore $u$ satisfies the difference equation

$$(5.18) \quad \prod_{j=1}^{s} (\Delta + \zeta_j I)^{m_j} u(t) = \sum_{k=1}^{p+1} \gamma_k \Delta^{k-1} u(t) = 0$$

which is of the form (5.3) with $d_k(t; \gamma) = \gamma_k$. This leads to

$$C_j^T = \Delta^{j-1}$$

and

$$X = \begin{pmatrix} c_1 & \cdots & c_1 \\ \vdots & \ddots & \vdots \\ c_p+1 & \cdots & c_{p+1} \end{pmatrix}$$
where \( c = U\gamma \) and \( U \) is defined by (5.13). Note that \( \zeta_j + \beta_j \) as \( n \to \infty \), so \( \gamma_k \to \zeta_k \).

Since \( \zeta \) is a linear transformation of \( \gamma \) for each \( n \), it can itself be used as a Prony parametrization. This leads to \( c_j^T = \mu^{j-1} \) and so on. Write \( \zeta \) for \( \zeta \) scaled to have elements which are \( O(1) \).

Scaling so that \( c_{p+1} = 1 \), the \( c_k \) satisfy

\[
\sum_{k=1}^{p+1} c_k \mu^{k-1}(t) = \prod_{j=1}^s \left( \prod_{i=1}^{m_j} (\mu_j - \rho_j) \right) \mu(t) = 0
\]

with \( \rho_j = e^{-\beta_j t} \). We will call \( \zeta \) the recurrence form parametrization, and \( \gamma \) the difference form. If the scales are chosen so that \( \gamma_{p+1} = c_{p+1} \), the difference form Prony matrix is related to be recurrence form matrix \( B_{\zeta} \) through

(5.19) \[
B(\gamma) = n^{-2p} U B_{\zeta} U.
\]

The recurrence form is not a suitable parametrization for asymptotic purposes since \( \rho_j \to 1 \) as \( n \to \infty \), so \( c_k \to (-1)^{p+1} \left( \frac{p}{k-1} \right) \), a limit that is independent of the parameters; we keep \( \zeta \) in a Roman font to emphasise this. Its advantage is that it is the simplest parametrization for which to calculate the Prony matrix. In the sequel it will always be the difference form to which we refer unless the recurrence form is explicitly mentioned.

Henceforth we will assume that the roots of \( p_{\xi}(\cdot) \) are distinct and real, so that the general solution for \( \mu(t) \) collapses to a sum of exponentials

\[
\mu(t) = \sum_{j=1}^{P} \alpha_j e^{-\beta_j t}.
\]

Then (if \( \xi_{p+1} = 1 \)) the \( \xi_k \) are the elementary symmetric functions.
of the $\beta_j$,

$$\xi_1 = \prod_{j=1}^{p} \beta_j$$

$$\xi_2 = \sum_{k=1}^{p} \prod_{j=1}^{p} \beta_j$$

$$\therefore$$

$$\xi_p = \sum_{k=1}^{p} \beta_j$$

$$\xi_{p+1} = 1$$

which we write as $\zeta = \operatorname{esf}(\beta)$. Similarly $\gamma = \operatorname{esf}(\zeta)$ if $\gamma_{p+1} = 1$, and $\zeta = \operatorname{esf}(\gamma)$ if $\zeta_{p+1} = 1$.

Historical remark. In a paper in the 1795 *Journal de l'Ecole Polytechnique* (see Kay and Marple, 1981), Prony considered the problem of interpolating a sum of $p$ exponentials through $2p$ points $y_i$. In effect, he used the recurrence from parametrization and solved the linear system

$$X(\zeta)^T \gamma = 0.$$ 

We attach his name to the more general class of parametrizations defined in §5.4.

5.9 Rational Fitting

Another class of linear homogeneous differential equations for which there are corresponding difference equations are

$$p_a(D)(q(t;\gamma)\mu(t)) = 0$$

where $p_a(\cdot)$ is any polynomial with known coefficients, and $q(t;\gamma)$
is any continuous function of time, linear and homogeneous in $\gamma$.

The general solution for $\mu$ has the form

$$\mu(t) = \frac{f(t)}{q(t;\gamma)}$$

where $f(t)$ is the general solution of $p_a(D)f(t) = 0$. In matrix terms this corresponds to

$$\mu = (Z\gamma)^{-1}W\alpha$$

where $W$ and $Z$ are known matrices of covariates.

Rational functions arise in the particular case that $p_a(D) = \varphi^P$ and $q(t;\gamma) = p_\gamma(t) = \sum_{j=1}^{q+1} \gamma_j t^{j-1}$. The general solution for $\mu$ is then

$$(5.20) \quad \mu(t) = \frac{p_\alpha(t)}{p_\gamma(t)}$$

with $p_\alpha(t) = \sum_{j=1}^q \alpha_j t^{j-1}$. We will also write $p_\gamma(t) = 1 + \sum_{j=1}^q \beta_j t^j$ when it is desired to have a parametrization of minimal dimension.

The rational function (5.20) satisfies the difference equation

$$\Delta^P(p_\gamma(t)\mu(t)) = 0$$

which is of the form (5.3) with $c_k(t;\gamma) = \Delta^{p-k+1}p_\gamma(t)$. This leads to

$$c^t = \Delta^P(p_\gamma(t)\mu)$$

and hence

$$c_j^t = \frac{\partial c^t}{\partial \gamma_j} = \Delta^P(t^{j-1}).$$

We will make the assumption throughout that $|p_\gamma(t)| > 0$ on $[0,1]$.

Remark. The Prony parametrization produces an interesting unification of the exponential and rational fitting examples. Consider the
expression (5.6) for $\mu$. In the exponential fitting case $C$ is a circulant matrix. Diagonalizing leads to

$$\mu = F(p_c(\omega))^{-1} F^T \alpha$$

where the components of $\omega$ are the $n$th roots of unity, and

$$p_c(z) = \sum_{k=1}^{p+1} c_k z^{-k}.$$ 

This displays $F^* \mu$ as a rational function of the complex frequency $\omega$. 
CHAPTER 6
A MODIFIED PRONY ALGORITHM

6.1 Introduction

In chapter 5 we showed that if \( \mu(t) \) satisfies a linear, homogeneous difference equation, expressed in matrix terms as

\[
X(\gamma)^T \mu = 0,
\]

then the least squares problem may be solved by minimizing

\[
\psi(\gamma) = \gamma^T P_X \gamma
\]

with \( P_X \) the projection onto \( R(X) \). Furthermore

\[
\dot{\psi} = B(\gamma) \gamma
\]

where \( B \) is a symmetric matrix function of \( \gamma \). Sections 6.2 and 6.3 describe an algorithm for solving \( B(\gamma) \gamma = 0 \), including some computational details, and section 6.4 examines the eigenstructure of \( B \). Section 6.5 demonstrates the inconsistency of a naive generalization of Prony's method. Section 6.6 deals with the application of the algorithm to exponential fitting, and section 6.7 with its modification in the presence of a linear parameter constraint.

6.2 A Modified Prony Algorithm

6.2.1 A sequence of linear eigenproblems

Henceforth we will scale \( \gamma \) to have norm one, this being the natural scale if \( \gamma \) is to be viewed as an eigenvector. We minimize
\( \psi(\gamma) \) subject to this constraint. Let
\[
F(\gamma, \lambda) = \psi(\gamma) + \lambda(1 - \gamma^T \gamma)
\]
where \( \lambda \) is a Lagrange multiplier. Then
\[
\dot{F}_\gamma = 2B(\gamma)\gamma - 2\lambda \gamma
\]
and
\[
\dot{F}_\lambda = 1 - \gamma^T \gamma
\]
so the necessary conditions for a minimum are
\[
(6.1) \quad (B(\gamma) - \lambda I)\gamma = 0, \quad \gamma^T \gamma = 1.
\]
Now \( \lambda \) must be zero at a solution of (6.1). Since \( \psi(\gamma) \) does not depend on \( \|\gamma\| \), \( \psi \) must be in a direction orthogonal to \( \gamma \), that is
\[
\gamma^T \psi(\gamma) = \gamma^T B(\gamma) \gamma = 0.
\]
Premultiplying \( \dot{F}_\gamma = 0 \) by \( \gamma^T \) gives \( \lambda = 0 \). This suggests the following iteration. Solve
\[
(6.2) \quad (B(\gamma^k) - \lambda^{k+1} I)\gamma^{k+1} = 0, \quad \gamma^{k+1 T} \gamma^{k+1} = 1
\]
with \( \lambda^{k+1} \) the nearest to zero of such solutions, and accept convergence when \( \lambda^{k+1} \) is small compared with \( \|B\| \). We call this the modified Prony algorithm.

6.2.2 Implementation

The linear eigenproblem (6.2) is solved extremely rapidly by inverse iteration (Wilkinson, 1965). And only an order of magnitude estimate of \( \|B\| \) is required: I have found \( \frac{1}{q+1} \sum |B_{ij}| \) to be adequate. My version of the modified Prony algorithm is
Algorithm 6.1

\[ \gamma^0 := \text{starting value} \]
\[ \lambda^0 := 0 \]
\[ k := 0 \]

\begin{align*}
\text{repeat} \quad & \\
\quad & \varepsilon := \tau \frac{1}{q+1} \sum_{i,j} |B_{ij}| \\
\quad & \gamma^0 := \gamma^k \\
\quad & \eta^0 := \lambda^k \\
\quad & l := 0 \\
\text{repeat (inverse iteration)} \quad & \\
\quad & w^{l+1} := (B - \eta^l I)^{-1} \gamma^l \\
\quad & \gamma^{l+1} := w^{l+1} / \|w^{l+1}\|_\infty \\
\quad & \eta^{l+2} := \eta^l + \frac{w^{l+1} \gamma^l + w^{l+2} \gamma^{l+2}}{\|w^{l+2}\|_\infty} \\
\quad & \gamma^{l+2} := w^{l+2} / \|w^{l+2}\|_2 \\
\quad & l := l+2 \\
\text{until} \quad & |\eta^l - \eta^{l-2}| < \varepsilon \\
\quad & \gamma^{k+1} := \gamma^l \\
\quad & \lambda^{k+1} := \eta^l \\
\quad & k := k+1 \\
\text{until} \quad & |\lambda^k| < \varepsilon
\end{align*}

The symmetric system

\[ (B - \eta^l I)w^{l+1} = \gamma^l \]
may be solved by Gaussian elimination, or by a method adapted to symmetric systems such as diagonal pivoting (Bunch and Kaufman, 1977).

The choice of the constant \( \tau \) (tolerance) reflects the precision required and the maximum condition number of a matrix that can be stored in finite arithmetic. For example, in double precision on a 36 bit machine, \( \tau = 10^{-15} \) would give precise solutions.

Empirical evidence suggests that the modified Prony algorithm has an impressive ability to converge to some stationary value even from quite poor starting values. As a general procedure for determining starting values, the parameter values for which \( \mu(t) \) is constant may be used. This gives \( y = s_1 \) in both the rational and exponential fitting examples (or \( c_k = (-1)^{p-k+1} \binom{p}{k-1} \) for the recurrence form parameters).

6.2.3 Choice of scale

Suppose we had imposed the quadratic constraint \( y^T A y = 1 \) instead of \( y^T y = 1 \). We would then have minimized

\[
F(\gamma, \lambda) = \psi(\gamma) + \lambda(1 - y^T A y)
\]

leading to the generalized nonlinear eigenproblem

\[
(B - \lambda A)\dot{y} = 0
\]

and to the sequence of linear problems

\[
(B^{(k)} - \lambda^{k+1} A)\dot{y}^{k+1} = 0.
\]

Now (6.3) is equivalent to

\[
(A^{-k} B^{(k)} A^{-kT} - \lambda^{k+1} I)\dot{\varrho}^{k+1} = 0
\]

with \( \dot{\varrho} = A^{kT} \dot{y} \). So using \( A \) as the inner product matrix is equivalent to the linear reparametrization \( \dot{\varrho} = A^{kT} \dot{y} \). Although the
modified Prony sequence for $\tilde{c}$ is different to that for $\gamma$, it has a similar convergence matrix, as we will show in the next chapter.

6.3 Calculation of $B$

6.3.1 The general case

Recall the expression (5.8) for $B$. The following scheme is suggested for computation.

$$Y := (X^T y \ldots X^T_{q+1} y)$$

A $(n-p) \times (q+1)$ matrix which does not depend on $\gamma$.

$$M := (X^T X)^{1/2}$$

The Choleski factor, a lower triangular $(n-p)^2$ matrix with $q+1$ bands.

$$Y := M^{-1} Y$$

$$(X^T X)^{-1/2} X^T_j y, j = 1, \ldots, q+1.$$  

$$Y := Y \gamma$$

$$(X^T X)^{-1/2} X^T \gamma, \text{ an } n-p \text{ vector.}$$

$$V := (X^T_{1-} \ldots X^T_{q+1-})$$

$$X_j (X^T X)^{-1/2} X^T \gamma, j = 1, \ldots, q+1; \text{ an } n \times (q+1) \text{ matrix.}$$

$$B := Y^T Y - V^T V.$$  

Flops required are $O(nq^2)$, for which the banded structure of $X$ is crucial. The fitted means may be extracted by

$$\hat{\gamma} = P_X \gamma = V \gamma.$$  

(6.4)

6.3.2 Exponential fitting

For exponential fitting we have $X^T_{j-} y = \Delta^{j-1} y$, which may be calculated recursively. Similarly for $X^T_{j-} v = \Delta^{j-1} T^T \gamma$. The matrix $X$ is toeplitz as well as banded, so only $q+1$ elements need to be stored. These are calculated from $\gamma = U \gamma$ with $U$ defined by (5.13).

For the recurrence form the scheme is even simpler. The elements
of $X$ are now the supplied parameters. The $X^T_{j-\infty}$ are simply windowed
shifts of $y$, and

$$(v^Tv)_{ij} = \sum_{k=1}^{n-p-|i-j|} v_k v_{k+i-j}.$$

The difference form Prony matrix can be calculated via the
recurrence form using (5.19) and (5.14). This is equivalent to

$$B_{ij} = n^{-2p(i-1)T \Delta^{-1}} B_{ij}.$$ 

Substantial subtractive cancellation must occur, though, because the
elements of $B$ are large and nearly equal. So it is necessary to
calculate $B$ directly.

6.3.3 Rational fitting

For rational fitting, the $X^T_{j-\infty} = P\Delta^p(t^{j-1})y$ are calculated by
repeated differencing of the $(t^{j-1})y$, and $X^T$ is $P\Delta^p(p_y(t))$. Rather than store $V$, calculate $w = \Delta^T P^T v$ and use

$$(v^Tv)_{ij} = w(t^{i+j-2})w.$$ 

For the more general, but closely related, problem of fitting

$$u = (Zy)^{-1} w \alpha$$

we would have $X_j = (z_i)H$, with $z_i$ the $i$th column of $Z$, and

$$H = \begin{bmatrix}
h_{11} & \cdots & \cdots \\
\vdots & \ddots & \vdots \\
h_{1,p+1} & \cdots & \cdots \\
h_{n-p,1} & \cdots & \cdots \\
h_{n-p,p+1} & \cdots & \cdots 
\end{bmatrix}.$$
determined so that \((h_{j,1}, \ldots, h_{j,p+1})^T\) is orthogonal to
\[
\begin{pmatrix}
  w_{j,1} & \cdots & w_{j,p} \\
  \vdots & & \vdots \\
  \vdots & & \vdots \\
  w_{j+p,1} & \cdots & w_{j+p,1}
\end{pmatrix}
\]

6.4 Eigenstructure of \(B\)

Lemma 6.1. *The least and greatest singular values of \(X\) are \(O(1)\) and \(O(n^P)\) respectively.*

Proof. The vectors which correspond to small singular values of \(X\) are obtained from realizations of smooth functions at the time points \(t_i\). Let \(f\) be a \(p\) times continuously differentiable function on \([0,1]\) that is not a solution of the differential equation (5.15).

Write
\[
\|f\|^2 = \int_0^1 f(t)^2 \, dt
\]
and let
\[
\tilde{z} = \|f(t)\|^{-1} f(t).
\]
Then
\[
(X^T \tilde{z})_{n(i)} = \|f(t)\|^{-1} \sum_{k=1}^{p+1} d_k(t_{n(i)}; \gamma) \Delta^k f(t_{n(i)})
\]
\[
+ \|f\|^{-1} \sum_{k=1}^{p+1} b_k(t_{s(i)}; \xi) D^k f(t_{s(i)})
\]
which is a constant.

On the other hand, the elements of \(X\) are \(O(n^P)\) through the contribution of \(\Delta^P\), so the largest singular value is \(O(n^P)\). \(\Box\)

Lemma 6.1 implies that the condition number of \(X\) will be large,
and appears to have consequences for the numerical accuracy of \( (X^TX)^{1/2} \), 
B and \( \hat{\mu} \). Experience with the exponential and rational fitting 
examples suggests, though, that this problem is mitigated by the fact 
that \( X \) possesses a compact analytic specification in terms of the 
Prony parameters.

The condition number of \( B \) itself is \( O(1) \), as is shown by the 
following theorem.

**Theorem 6.1.** For exponential and rational fitting, \( n^{-1}B(\hat{\gamma}) \) is 
asymptotically equal to \( n^{-1}\mu_P^T X\hat{u} \) with probability one.

The proof of this theorem is postponed until chapter 7, since it 
is long and closely related to the proof of stability. For the 
general case, not restricting to exponential or rational functions, I 
believe the theorem to remain true, but it is then a conjecture. A 
method by which the proof might be extended is briefly mentioned in 
chapter 7.

**Corollary 6.1.** With probability one, \( n^{-1}B(\hat{\gamma}) \) has a positive semi-
definite limit, the Moore-Penrose inverse of which is the asymptotic 
covariance matrix of \( n\hat{\gamma} \).

**Corollary 6.2.** With probability one, the zero eigenvalue of \( B(\hat{\gamma}) \) is 
asymptotically isolated with multiplicity one.

### 6.5 Prony's Method

Prony's classical method consisted, in the context of exponential 
fitting, of solving the linear system

\[
X^T \hat{y} = 0
\]

for \( n = 2p \). A direct generalization to the overdetermined case is
to minimize

\[(6.5) \quad y^T X^T y. \]

The methods of Benson (1979) and Varah (1982) are closely related to this. Varah proposed minimizing a differential version of (6.5), after fitting a cubic spline through the \( y_i \). It is worth demonstrating therefore that the estimate which minimizes (6.5) is not consistent. The basic reason is that \( X^T y \) has a highly nontrivial covariance structure (lemma 6.1) that cannot be ignored.

Let

\[ Y = (X^T_{1,\ldots,q+1} y) . \]

Then

\[ y^T X^T y = y^T Y^T y, \]

so minimizing (6.5) is equivalent to finding the least eigenvalue of \( Y^T Y \). This matrix may be expanded as

\[(6.6) \quad (Y^T Y)_{ij} = y^T_{1,\ldots,q+1} y_{1,\ldots,q+1} \]

\[ = \mu_{1,\ldots,q+1} + \mu_{1,\ldots,q+1} + \varepsilon_{1,\ldots,q+1} + \varepsilon_{1,\ldots,q+1} + \varepsilon_{1,\ldots,q+1}. \]

Let \( k-1 \) be the highest power of \( \Delta \) which contributes to \( X_{x,1} \), and similarly \( l-1 \) for \( X_{x,1} \), then

\[ \mathbb{E} \left[ \frac{1}{n} \varepsilon_{1,\ldots,q+1} \right] = \frac{\sigma^2}{n} \text{tr}(X_{1,1}^T X_{1,1}) = o(n^{k+l-2}) . \]

On the other hand, \( n^{-1} \mu_{1,\ldots,q+1} \) has a finite limit, so it is clear that the systematic component of \( Y^T Y \) is swamped by random components.

6.6 More on Exponential Fitting

6.6.1 Consistency

Recall that the solution of the difference equation (5.18) is a
sum of exponentials only if the polynomial \( p_y(x) \) has distinct real roots. In other words, the restriction of \( y \) to be real is a richer parametrization than that of \( \beta \) to be real. If fitting a sum of exponentials is our primary problem, then a question of consistency arises. On the face of it, it is possible that the Prony parameter vector \( y \), obtained from the elementary symmetric functions of the least squares estimate \( \hat{\beta} \) of \( \beta \), does not solve \( \hat{\psi} = 0 \). We show now that this is not the case: there is always a solution of \( \hat{\psi} = 0 \) corresponding to \( \hat{\beta} \), even though it may not be a global minimum of \( \psi(y) \).

Lemma 6.2. Let \( s = esf(\zeta) \). If the \( \zeta_j \) are distinct, then \( s_\zeta \) has full column rank.

Proof. The elementary symmetric functions obey a recurrence relationship which is neatly expressed by

\[
s = \prod_{j=1}^{p} (\zeta_j I + \Pi^T) e_1 .
\]

This allows the proof to proceed by induction. Note firstly that the result is true for \( p = 2 \), since then

\[
\mathbf{s}_\zeta = \begin{bmatrix} \zeta_2 & \zeta_1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}
\]

which has rank 2. Let

\[
r = \prod_{j=1}^{p-1} (\zeta_j I + \Pi^T) e_1 .
\]

Then \( r_{p+1} = 0 \), and \( r_1, \ldots, r_p \) are the elementary symmetric functions of \( \zeta_1, \ldots, \zeta_{p-1} \). Suppose
\[
\dot{s}_u = 0
\]

and expand the left hand side as

\[
\dot{s}_u = \sum_{j=1}^{p-1} \dot{s}_u \hat{u}_j + \dot{s}_u
\]

\[
= (\zeta P^I + \Pi^T) \sum_{j=1}^{p-1} \dot{r}_j \hat{u}_j + ru
\]

where \( \dot{s}_j = \frac{\partial s}{\partial \zeta_j} \) and \( \dot{r}_j = \frac{\partial r}{\partial \zeta_j} \). Suppose that the \( \zeta_j \) have been permuted so that \( \zeta_P \) is not on the unit circle. Then \( (\zeta P^I + \Pi^T) \) is nonsingular, and

\[
\sum_{j=1}^{p-1} \dot{r}_j \hat{u}_j + (\zeta P^I + \Pi^T) \hat{u}_P = 0 .
\]

Our approach will be to show that at least one of the last two elements of \( (\zeta P^I + \Pi^T)^{-1} \hat{r} \) is nonzero. This implies that \( u_P = 0 \) since the last two elements of each \( \hat{r}_j \) are zero, hence

\[
\sum_{j=1}^{p-1} \dot{r}_j \hat{u}_j = 0
\]

and \( u = 0 \) by the induction hypothesis.

If \( \zeta_P = 0 \) then

\[
(\zeta P^I + \Pi^T)^{-1} \hat{r} = \Pi \hat{r}_1 ,
\]

the last element of which is

\[
r_1 = \prod_{j=1}^{p-1} \zeta_j .
\]

This must be nonzero since each of the \( \zeta_j \)'s in the product must be.

If \( \zeta_P \) is nonzero, then expanding \( (I + \zeta P^{-1} \Pi^T)^{-1} \) as a geometric series, and using \( \Pi P^{T+1} = I \), gives
\[(\zeta_p I + \Pi^T)^{-1} = -\zeta_p (1 - (-\zeta_p)^{p+1})^{-1} \sum_{k=1}^{p+1} (-\zeta_p)^{p-k} \Pi^T \zeta^{-1} .\]

The pth element of
\[\sum_{k=1}^{p+1} (-\zeta_p)^{k-1} \zeta_k = \prod_{j=1}^{p-1} (\zeta_j - \zeta_p)\]
is nonzero since the \(\zeta_j\) are distinct.

\textbf{Lemma 6.3.} If the \(\beta_j\) are distinct, then \(\dot{\gamma}_\beta\) has full column rank.

\textbf{Proof.} Recall that \(\gamma = \Pi_s^{-1} s\) with \(s = \text{esf}(\beta)\) and \(\zeta = n(1 - e^{-\beta/n})\). Therefore
\[\dot{\gamma}_\beta = \dot{s} s^T \zeta \beta\cdot\]
Suppose that
\[(6.7) \quad \dot{s} s^T \zeta \beta u = 0 .\]
Now
\[\dot{s} = \Pi_s^{-1} (I - \gamma \gamma^T)\]
which has the rank \(p\), and null space spanned by \(\gamma\). And \(\gamma\) does not belong to the range space of \(\dot{s} \zeta\) since \(\gamma_{p+1} = \Pi_s^{-1}\) while the \(p+1\)th row of \(\dot{s} \zeta\) is zero. So (6.7) implies
\[\dot{s} s^T \zeta \beta^\top u = 0 ,\]
and hence \(u = 0\) by lemma 6.2 and the fact that
\[\dot{\zeta}_\beta = (e^{-\beta/n})\]
has full rank.

\textbf{Corollary 6.3.} The null space of \(\dot{\gamma}_\beta^\top\) is spanned by \(\gamma\).
Proof. That $\gamma$ belongs to the null space follows from differentiating $\gamma^T Y = 1$. It must then span the space, which has dimension one. \hfill $\square$

Theorem 6.2. If $\beta$ solves $\dot{\gamma}_\beta = 0$, and the $\beta_j$ are distinct, then $\gamma(\beta)$ solves $\dot{\gamma}_Y = 0$.

Proof. By the chain rule

$$\dot{\gamma}_\beta = \gamma^T \dot{\gamma}_Y,$$

which can be modified to

$$(\dot{\gamma}_\beta, 0)^T = (\gamma^T \gamma)^T \dot{\gamma}_Y$$

since $\gamma^T \dot{\gamma}_Y = 0$. Corollary 6.3 implies that $(\gamma^T \gamma)$ is nonsingular, so

$$\dot{\gamma}_Y = (\gamma^T \gamma)^{-T}(\dot{\gamma}_\beta, 0)^T.$$

\hfill $\square$

Remark. Whenever exponential fitting is discussed in this thesis, it is assumed that estimation of the rate constants $\beta_j$ is the primary problem. In practice though, it is usually the case that the exponential functions arose in the first place as solutions of a differential equation. This being so, it would seem more natural to take the estimation of the differential equation to be the primary problem, and to minimize $\phi$ with respect to $\gamma$ even if this results in $\rho_Y(\cdot)$ having complex or multiple roots.

6.6.2 Recovery of the rate constants

Having estimated $\gamma$, we can obtain $\gamma$ directly from (6.4). But usually we will want to recover the rate constants $\beta_j$ for the purpose of interpretation. In the recurrence form we must solve

$$P_c(z) = 0$$
to obtain roots $\rho_j = e^{-\beta_j/n}$. For large $n$ this is an ill-conditioned problem because the $\rho_j$ cluster near 1. Another aspect of the same problem is that asymptotically the leading significant figures of the $\zeta_k$ contain no information about the $\beta_j$.

This problem does not arise in the difference formulation: we solve

$$p_\gamma(z) = 0$$

to obtain roots

$$\zeta_j = n(1 - e^{-\beta_j/n}).$$

The final step

$$\beta_j = -n \log(1 - \zeta_j/n)$$

$$= n \sum_{j=1}^{\infty} j^{-1}(\zeta_j/n)^j$$

will cause problems only if $\zeta_j$ is large and negative.

6.6.3 Eigenstructure of $B_c$

The recurrence form parametrization depends on $n$ in an essential way: powers of $n$ enter in a way which dominates the eigenstructure of $B_c$.

Theorem 6.3. Let $\lambda_1, \ldots, \lambda_p$ be the eigenvalues of $B_c(\hat{\gamma})$ in ascending order. Then

$$\lambda_k = O(n^{2k-1}) \quad k = 2, \ldots, p+1.$$ 

Proof. Let

$$v_{ij} = (-1)^{i-j} \begin{bmatrix} i-1 \\ j-1 \end{bmatrix}$$

so that
From (5.19) we have that

\[ U = V \begin{bmatrix} 1 & \cdots & n \end{bmatrix} \]

Let \( f_k \) be a polynomial of degree \( k-1 \) which satisfies \( f_k(i) = 0 \), for \( i = 1, \ldots, k-1 \), and \( f_k(k) = 1 \). Let \( \mathbf{v}_k = (\mathbf{f}_k(1) \ldots \mathbf{f}_k(p+1))^T \).

Then

\[ \mathbf{v}_k^T \mathbf{v}_k = e_k \]

so the \( \mathbf{v}_k \) are the columns of \( \mathbf{V}^T \), and

\[ B_c = V^{-T} \begin{bmatrix} n^p \\ \vdots \\ n \\ l \end{bmatrix} \begin{bmatrix} n^p \\ \vdots \\ n \\ l \end{bmatrix} \mathbf{V}^{-1} \]

Now, for large \( n \), all proper submatrices of \( B(\hat{\gamma}) \) are nonsingular. This follows because \( \hat{\gamma}_1 \) and \( \hat{\gamma}_{p+1} \) are nonzero (none of the true rate constants \( \beta_{oj} \) may be zero) and \( \hat{\gamma} \) spans the null space of \( B(\hat{\gamma}) \). In particular, the diagonal elements of \( B(\hat{\gamma}) \) are nonzero. From Corollary 6.1 they are \( O(n) \).

Let \( x_1, \ldots, x_{p+1} \) be the orthonormal sequence obtained from \( \mathbf{v}_1, \ldots, \mathbf{v}_{p+1} \) by Gram-Schmidt orthonormalization. This is equivalent to
where \((VV^T)^{1/2}\) is the Choleski factor of \(VV^T\). The largest eigenvalue is

\[
\lambda_{p+1} = \max_{z \sim \mathcal{C}_\sim} z^T B z.
\]

Asymptotically, this is achieved by \(z = \hat{x}_1 = (p+1)^{-1/2}\), and is \(O(n^{2(p+1) - 1})\). Defining the eigenvalues recursively, the \(k\)th largest eigenvalue is asymptotically equal to (more strictly, bounded above by)

\[
\max_{z \sim \mathcal{C}_\sim} \max_{z \sim \mathcal{C}_\sim} z^T B z
\]

which is asymptotically achieved by \(z = \hat{x}_k\), and is \(O(n^{2k-1})\).

6.7 Modified Prony with Linear Constraints

6.7.1 Augmenting the Prony matrix

Suppose we have prior information about \(\gamma\) which can be expressed as a linear constraint \(g^T \gamma = 0\). We then minimize

\[
P(\gamma, \lambda, \eta) = \psi(\gamma) + \lambda (1 - \gamma^T \gamma) + 2\eta s \gamma^T g
\]

where \(\lambda\) and \(\eta\) are Lagrange multipliers; \(s\) is a scale factor chosen to facilitate convergence of the inverse iteration sequence.

Differentiating gives

\[
\begin{align*}
\dot{P}_\gamma &= 2B(\gamma) \gamma - 2\lambda \gamma + 2\eta sg \\
\dot{P}_\lambda &= 1 - \gamma^T \gamma \\
\dot{P}_\eta &= 2sg \gamma^T
\end{align*}
\]

Let
Then the necessary conditions for a minimum may be summarized as

\[(A - \lambda P)v = 0, \quad v^T P v = 1\]

which is a generalized eigenproblem. We again find, by premultiplying \( \hat{F} \gamma = 0 \) by \( \gamma^T \), that \( \lambda = 0 \) at a solution. We solve (6.8) by solving the sequence of linear problems

\[(A(y^k) - \lambda^{k+1} P)v^{k+1} = 0, \quad v^T v^{k+1} = 0.\]

The inverse iteration sequence of Algorithm 6.1 is modified to

repeat (inverse iteration)

\[w^{l+1} = (A - \eta^l P)^{-1} v^l,\]
\[v^{l+1} = \frac{w^{l+1}}{\|w^{l+1}\|} v^l,\]
\[w^{l+2} = (A - \eta^l P)^{-1} w^{l+1},\]
\[\eta^{l+2} = \eta^l + \frac{w^{l+2}}{\|w^{l+2}\|} v^{l+1},\]
\[v^{l+2} = \frac{w^{l+2}}{\|w^{l+2}\|} P w^{l+2} v^l,\]
\[\lambda = \lambda^{l+2}\]

until \( |\eta^l - \eta^{l-2}| < \varepsilon \).

The eigenvalues of (6.9) are unaffected by \( s \), since

\[
\det(A - \lambda P) = \det \begin{bmatrix} B - \lambda I & sg \\ sg^T & 0 \end{bmatrix} = s^2 \det \begin{bmatrix} B - \lambda I & g \\ g^T & 0 \end{bmatrix}.
\]

So we can take \( s = 1 \) without affecting the rate of convergence of
the inverse iteration sequence. And this determinant is a polynomial in $\lambda$ of order only $q$. So the dimension of the eigenproblem has been reduced by the constraint, even though the dimension of the Prony matrix has increased.

6.7.2 Deflating the Prony matrix

An alternative approach to the constraint is to explicitly deflate the dimension of $B$. Let $Q$ be the projection onto $R(q)^\perp$, 

$$Q = I - (q^T q)^{-1} q q^T.$$ 

Premultiplying $\hat{F}$ by $Q$ allows us to express the necessary conditions for a minimum as 

$$\begin{align*}
(QB - \lambda I) y &= 0, \quad y \in R(Q), \\
\gamma^T y &= 1.
\end{align*}$$

Let $W$ be a $(p+1) \times p$ matrix spanning $R(Q)$, so 

$$Q = W(W^T W)^{-1} W^T.$$ 

Let $W_\delta = \gamma$. Then (6.10) may be re-expressed as 

$$\begin{align*}
(W^T B W - \lambda W^T W) \delta &= 0, \\
\delta^T W_\delta &= 1.
\end{align*}$$

Solving the linear problem 

$$\begin{align*}
(W^T B (\gamma^T W - \lambda W^T W) \delta &= 0, \\
\delta^T W_\delta &= 1
\end{align*}$$

and recovering $\gamma^{k+1}$ as $W_\delta^{k+1}$, is equivalent to solving (6.9).

The eigenvalues of (6.11) are the stationary values of 

$$\begin{align*}
\frac{T W^T B W z}{z^T W^T W z} \\
\frac{T W^T z}{z^T W^T z}
\end{align*}$$

which are those of 

$$\begin{align*}
\frac{T B y}{y^T T y} \\
\frac{T y}{y^T y}
\end{align*}$$
subject to the constraint that \( \mathbf{v}^T \mathbf{g} = 0 \). Hence, from the maximum
minimum characterization of eigenvalues, the eigenvalues of (6.11)
interlace those of \( B(y_k) \). Let \( \nu_1, \ldots, \nu_q \) be the eigenvalues of
(6.11), and \( \lambda_1, \ldots, \lambda_{q+1} \) be those of \( B \). Then
\[
\lambda_1 \leq \nu_1 \leq \ldots \leq \nu_q \leq \lambda_{q+1}.
\]

A possible general approach to the calculation of \( \mathbf{W} \) is the
following. Determine the unit vector \( \mathbf{w} \) such that \( \mathbf{g} \) is the first
column of the orthogonal matrix
\[
\mathbf{I} = 2\mathbf{w}\mathbf{w}^T
\]
and take \( \mathbf{W} \) to be the other \( q \) columns. Then \( \mathbf{W}^T \mathbf{W} = \mathbf{I} \) and \( \mathbf{W} \mathbf{B} \mathbf{W} \)
is the trailing \( q \times q \) submatrix of
\[
(\mathbf{I} - 2\mathbf{w}\mathbf{w}^T) \mathbf{B} (\mathbf{I} - 2\mathbf{w}\mathbf{w}^T).
\]

6.7.3 Exponential fitting

Suppose the mean function includes a constant term, say
\[
\mu(t) = \alpha_1 + \sum_{j=2}^{p} \alpha_j e^{-\beta_j t}.
\]
That is, we know \( \beta_1 = 0 \). This is equivalent to the linear
constraint \( e_1^T \mathbf{y} = 0 \). The projection onto \( \mathbb{R}(e_1)^\perp \) is \( \mathbf{W} \mathbf{W}^T \) with
\( \mathbf{W}^T = (0 \ I_q) \), so we can solve
\[
(\mathbf{W} \mathbf{B}(y_k)^k - \lambda_k^{k+1} \mathbf{I}) \mathbf{y}_k^{k+1} = 0, \quad \mathbf{y}_k^{k+1} = 1,
\]
and recover \( \mathbf{y}_k^{k+1} \) as \( (0 \mathbf{A}_{k+1}^T)^T \). Here \( \mathbf{W} \mathbf{B} \mathbf{W} \) is simply the
trailing \( q \times q \) submatrix of \( \mathbf{B} \).

In the recurrence formulation, the constraint \( e_1^T \mathbf{y} = 0 \)
transforms to \( \mathbf{1}^T \mathbf{c} = 0 \). A suitable deflated matrix may be calculated
using the orthogonal matrix $I - 2\omega \omega^T$ with

\[
\omega_1 = \left( \frac{4p+4}{5p+4} \right)^{\frac{1}{2}}
\]

and

\[
\omega_j = \left( \frac{8(p+1)(p+2)}{5p+4} \right)^{\frac{1}{2}} \quad j = 2, \ldots, p+1.
\]

But in this formulation it seems simpler to solve the constrained problem by augmenting rather than deflating the Prony matrix.
CHAPTER 7
ASYMPTOTIC STABILITY

7.1 Introduction

In this chapter we prove the asymptotic stability of the modified Prony iteration for exponential and rational fitting. In sections 7.2 and 7.3 we show that the convergence matrix may be expressed as

\[ B^+B^\top \gamma \]

and that the spectrum of this matrix is the same for all Prony parametrizations. Section 7.4 gives expectations for \( B(\gamma_0) \) and \( \hat{B}(\gamma_0)\gamma_0 \). Sections 7.5 to 7.7 prove the stability result, that

\[ \rho(B^+(\gamma)\hat{B}(\gamma)\gamma) \xrightarrow{a.s} 0. \]

Only in the last two sections, which prove the convergence of \( n^{-1}B(\gamma) \) itself, is resort made to special features of the exponential and rational fitting examples.

7.2 The Convergence Matrix

Let us express the modified Prony iteration formally as

\[ \gamma^{k+1} = F(\gamma^k) \]

where \( F \) is defined implicitly by (6.2). The iteration is stable if

\[ \rho(F(\gamma)) < 1. \]

Differentiating (6.2) with respect to \( \gamma^k \), and dropping the iteration count superscripts, gives
Here $\dot{\gamma}$ is a synonym for $\dot{\mathbf{P}}$. Now both $\lambda$ and $\dot{\lambda}$ are zero at $\hat{\gamma}$.

The latter is proved by premultiplying (7.1) by $\hat{\gamma}^T$, and by differentiating

$$\hat{\gamma}^T B(\gamma) \dot{\gamma} = 0$$

to show that

$$\hat{\gamma}^T B(\hat{\gamma}) \dot{\hat{\gamma}} = 0.$$ 

Therefore, evaluating (7.1) at $\gamma^k = \gamma^{k+1} = \hat{\gamma}$ gives

$$\dot{\gamma} + B \gamma = 0.$$ 

This may be solved for $\dot{\gamma}$ using

$$\dot{\gamma} = (B + \hat{\gamma} \hat{\gamma}^T) \gamma$$

so

$$\dot{\mathbf{P}}(\hat{\gamma}) = \dot{\gamma}(\hat{\gamma}) = (B + \hat{\gamma} \hat{\gamma}^T)^{-1} B \dot{\gamma}$$

$$= B \hat{\gamma}^T B(\hat{\gamma}) \dot{\hat{\gamma}}.$$ 

The condition for stability may be interpreted as being that $B$ should not change too rapidly as a function of $\gamma$. More precisely, the eigenvector associated with its smallest eigenvalue should not change too rapidly.

7.3 Invariance of the Convergence Factor

The following theorem shows that all Prony parametrizations, being related by linear transformations, have the same convergence properties.

Theorem 7.1. The eigenvalues of $B \gamma$ are invariant under linear reparametrizations.

Proof. Let $\delta$ be a parametrization related to $\gamma$ by $U \delta = \gamma$, where
U is a nonsingular known matrix. Then
\[ B_\delta = U^T B Y U \]
so
\[ \dot{B}_\delta = U^T (\dot{B} Y \dot{\delta}) U \]
\[ = U^T (\dot{B} Y) U . \]

Therefore
\[ B_\delta B_\delta = (U^T B Y U)^T U^T (\dot{B} Y) U \]
\[ = U^{-1} \dot{B} Y U \]
which is similar to
\[ B_\delta B_\delta Y \]

Theorem 7.1 holds for every n if the reparametrization depends on n. In particular, the recurrence form algorithm for exponential fitting converges if and only if the difference form does.

7.4 Expectations

Calculating expectations for B and \( \dot{B} Y \) supports the proposition that the convergence factor is eventually small. All expressions in this section are taken to be evaluated at the true parameter values.

Theorem 7.2.
\[ \mathbb{E}(B(Y_0)) = (y^T \lambda (X^T X)^{-1} X^T Y) \]
and
\[ \mathbb{E}(\dot{B}(Y_0) Y_0) = 0 . \]

Proof. Using \( X^T Y = X^T (Y - \mu) \) and trace identities, we write
\[ B_{ij} = \text{tr}(X_i^T (X^T X)^{-1} X_j^T Y Y^T) - \text{tr}(X^T X)^{-1} x_i^T X_j^T (X^T X)^{-1} Y^T (X^T X)^{-1} Y^T (X^T X)^{-1} (X^T X)^{-1} (Y - \mu)^T (Y - \mu)^T \]
which has expectation
\[
\text{tr}\{X_i(X^TX)^{-1}X_j^T(I\sigma^2 + \mu\mu^T)\} - \text{tr}\{X(X^TX)^{-1}X_j^T (X^TX)^{-1}X_i^T I\sigma^2\}
\]
\[= \mu_i^T X_i^T (X^TX)^{-1} X_j^T \mu_j .
\]

To do the same for \( \dot{\hat{\psi}}_\hat{\gamma} \) we need an explicit expression. The simplest way is to differentiate

\[\dot{\psi} = y^T P X_y^T \theta y \]

and to use

\[\ddot{\psi} = 2(B + \dot{B}_\gamma) .\]

This gives

\[(\dot{\hat{\psi}})_ij = - y_i^T X_i^T (X^TX)^{-1} X_j^T X (X^TX)^{-1} X^T \theta y_j - \text{symmetry term}
- y_i^T X_i^T (X^TX)^{-1} X^T X_j^T X (X^TX)^{-1} X^T \theta y_j - \text{symmetry term}
+ y_i^T X (X^TX)^{-1} X_j^T X (X^TX)^{-1} X_i^T X (X^TX)^{-1} X^T \theta y_j + \text{symmetry term}
+ y_i^T X (X^TX)^{-1} X_i^T X (X^TX)^{-1} X_j^T X (X^TX)^{-1} X^T \theta y_j
+ y_i^T X (X^TX)^{-1} X^T X_i^T (X^TX)^{-1} X_j^T X (X^TX)^{-1} X^T \theta y_j .
\]

By "symmetry term" we mean the one with \( i \) and \( j \) interchanged. The theorem follows from writing \( X^T \theta y = X^T (\theta y - \theta) \), applying the trace identities, and cancelling out like terms. \( \square \)

Note that

\[\dot{\mu}_\gamma^T X_i^T (X^TX)^{-1} X_j^T \mu_j = (\dot{\mu}_\gamma P X \mu)_ij
\]
since \( X^T \mu = 0 \) implies

\[X^T \dot{\mu}_\gamma = -X^T \mu_\gamma .
\]

In \( B \) and \( \dot{\hat{\psi}}_\gamma \) we have a decomposition of \( \dot{\psi} \) which is analogous to that of \( \Phi_\theta \) into \( \dot{\mu}_\theta^T \mu_\theta \) and \( -\dot{\mu}_\theta^T (y - \mu) \). Compare with the Gauss-Newton convergence matrix (1.15).
7.5 Stability

We prove the asymptotic stability of the modified Prony algorithm using Theorem 6.1 and the asymptotic convergence of \( n^{-1/2} \hat{\Psi}(\hat{\beta}) \). Recall that \( \gamma + \xi \), where \( \xi \) parametrizes the differential equation formulation (5.15). We will assume that there exists a reparametrization from \( \xi \) on the unit ball in \( \mathbb{R}^{q+1} \) to \( \beta \) in \( \mathbb{R}^{q} \), such that \( \gamma_\beta \) has full column rank in a neighbourhood of the true value \( \beta_0 \). That this is so for exponential fitting was shown in §6.2.1. It is easily verified for rational fitting with

\[
\beta^T = (\gamma_2 \cdots \gamma_{q+1}) / \gamma_1.
\]

Recall also from §5.5 that \( \mu = A \alpha \) with \( A \) the trailing \( p \) columns of \( C^{-T} \). The matrix \( A \) may be considered a function of \( \beta \), and the results and assumptions of §5.2 can be applied to \( \mu(t; \theta) \) with \( \theta^T = (\alpha^T, \beta^T) \). In particular

\[
\lim_{n \to \infty} \frac{1}{n} \dot{\phi}_\theta(\hat{\beta}) \overset{a.s.}{=} \lim_{n \to \infty} \frac{2}{n} \dot{\mu}_\theta(\theta_0)^T \dot{\mu}_\theta(\theta_0)
\]

which is positive definite. So

\[
\frac{1}{n} \dot{\psi}_\beta(\hat{\beta}) = \frac{1}{n} (\dot{\phi}_\beta - \dot{\phi}_\beta \phi_\alpha \phi_\beta^{-1} \phi_\alpha) (\hat{\alpha}, \hat{\beta})
\]

almost surely has the same limit as

\[
\frac{2}{n} \dot{\mu}_\beta(I - P_A \beta) \beta(\theta_0) = \frac{2}{n} \beta P x \dot{\mu}_\beta(\theta_0)
\]

which is positive definite. Therefore

\[
\frac{1}{n} \dot{\psi}_\gamma(\hat{\gamma}) = \frac{1}{n} (\dot{\gamma}_\beta(\hat{\gamma}) \dot{\gamma})^{-T} \begin{bmatrix} \dot{\psi}_\beta(\hat{\beta}) & 0 \\ 0 & 0 \end{bmatrix} (\dot{\gamma}_\beta(\hat{\gamma}) \dot{\gamma})^{-1}
\]

almost surely has the same limit as
which is positive semi-definite with null space spanned by $\gamma_0$. We will write

$$V_0 = \lim_{n \to \infty} \frac{1}{n} \mu_\gamma \mu_\gamma^T \gamma_0.$$

Lemma 7.1.

$$\frac{1}{n} (B(\gamma) + \hat{B}(\gamma) \gamma) \overset{a.s.}{\rightarrow} V_0.$$

Proof. Follows from

$$\hat{\psi} = 2(B + \hat{B} \gamma)$$

and the above discussion. □

Theorem 7.3. The modified Prony iteration applied to exponential or rational fitting is asymptotically numerically stable.

Proof. Theorem 6.1 (which is proved in the next section) is equivalent to

$$\frac{1}{n} B(\gamma) \overset{a.s.}{\rightarrow} V_0.$$

Combined with lemma 7.1, this proves

$$\frac{1}{n} \hat{B}(\gamma) \gamma \overset{a.s.}{\rightarrow} 0.$$

The convergence matrix may be expressed as the matrix ratio

$$\left(\frac{1}{n} B(\gamma) + \gamma \gamma^T\right)^{-1} \frac{1}{n} \hat{B}(\gamma) \gamma,$$

the spectral radius of which converges to zero, since the denominator converges to the positive definite matrix $V_0 + \gamma_0 \gamma_0^T$. □

7.6 Convergence of $B$

We return to theorem 6.1, which we prove in the form
We establish that $B(y)$ converges almost surely, uniformly in a neighbourhood of $y_0$, to a continuous function $V(y)$ which is equal to $V_0$ at $y_0$. It will be convenient to write

$$D_k = C_k C_k^{-1}, \quad k = 1, \ldots, q+1.$$ 

In the exponential fitting case, the $D_k$ are circulant matrices. In the rational case they are the diagonal matrices

$$\langle p_y(t)^{-1} t^{k-1} \rangle.$$ 

There are two crucial steps in the proof. The first is to express $B$ in terms of the $D_k$ and the projection $P_A$. Let $a_i$ be a column of $A$. The second step is to prove that each $D_k a_i$ and $D_j^T D_k a_i$ converges to a continuous function evaluated at $t$. Appeal can then be made to theorem 5.1 (law of large numbers) to prove results such as

$$\frac{1}{n} a_i^T D_x^T a_s \xrightarrow{a.s.} 0.$$ 

The following lemmas embody this second step. They are proved only for the exponential and rational fitting special cases. In fact the proofs are postponed until the next section because the exponential fitting case is quite tedious.

Lemma 7.2. For each $u$, there exist functions $f_j$, continuous on $[0,1]$, such that

$$(D_u A)_{ij} = f_j(t_i) + O\left(\frac{1}{n}\right)$$

uniformly for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

Lemma 7.3. For each $u$ and $v$, there exist functions $g_j$, continuous on $[0,1]$, such that
\[(D_{v}^{T}A_{0})_{ij} = g_{0j}(t_{i}) + O\left(\frac{1}{n}\right)\]

uniformly for \(i = 1, \ldots, n\) and \(j = 1, \ldots, p\).

Lemma 7.4. Let \(A_{0} = A(y_{0})\). For each \(u\) and \(v\), there exist functions \(g_{0j}\), continuous on \([0,1]\), such that

\[(D_{v}^{T}A_{0})_{ij} = g_{0j}(t_{i}) + O\left(\frac{1}{n}\right)\]

uniformly for \(i = 1, \ldots, n\) and \(j = 1, \ldots, p\).

Proof of theorem 6.1. Let \(P = (I_{n-p})\). Substituting

\[X_{1}^{T} = PC_{1}^{T} = PC^{T}C_{1}^{T} = X^{T}D_{1}^{T}\]

into the expression for \(B\) gives

\[B_{ij} = (D_{1}^{T}P_{1}D_{1}^{T}y - D_{1}^{T}P_{1}D_{1}^{T}y).\]

Expanding \(P_{1}\) as \(I - P_{A}\) and \(y\) as \(y_{0} + \varepsilon\) gives

\[\frac{1}{n}B_{ij} = \frac{1}{n} (y_{0} + \varepsilon)^{T}D_{i}(I - P_{A})D_{j}(y_{0} + \varepsilon) - \frac{1}{n} (y_{0} + \varepsilon)^{T}D_{i}(I - P_{A})D_{j}(I - P_{A})\]

\[\times (y_{0} + \varepsilon).\]

The terms

\[\frac{1}{n} \varepsilon D_{i}D_{j}^{T} = \frac{1}{n} \varepsilon D_{i}D_{j}^{T}\]

cancel out of this expansion, since \(D_{i}\) and \(D_{j}^{T}\) commute. Repeated application of lemmas 7.2 to 7.4 and theorem 5.1 shows that all other terms which involve \(\varepsilon\) converge to zero. We consider three terms as examples. Let

\[D_{i}A = F_{i} + E\]

\[D_{j}^{T}D_{i}A = G_{ij} + E\]

and

\[D_{j}^{T}D_{i}A_{0} = G_{0ij} + E\]
where the columns of $F_i$, $G_{ij}$ and $G_{0ij}$ are obtained by evaluating continuous functions at $\sim$, and $E$ represents any $n \times p$ matrix whose elements are uniformly $O\left(\frac{1}{n}\right)$. Thus for example

$$\frac{1}{n} D_{i,j}^T (F_i + E) = \frac{1}{n} D_{i,j}^T \left(\frac{1}{n} A^T A\right)^{-1} \frac{1}{n} (F_i + E)^T$$

converges almost surely to zero because $\frac{1}{n} D_{i,j}^T (F_i + E)$ converges to a constant $p$-vector, $\frac{1}{n} A^T A$ to a constant positive definite $p \times p$ matrix, and the $p$ elements of $\frac{1}{n} (F_i + E)^T$ converge almost surely to zero by theorem 5.1. Moreover the convergence is uniform for $\gamma$ in a compact set. Similarly

$$\frac{1}{n} D_{i,j}^T (G_{ij} + E) = \frac{1}{n} G_{ij} + E$$

and

$$\frac{1}{n} D_{i,j}^T (G_{0ij} + E) = \frac{1}{n} G_{0ij} + E$$

converge to zero. The other terms involving $E$ are treated in the same way. The only terms in the expansion for $\frac{1}{n} B_{ij}$ which remain are

$$\frac{1}{n} D_{i,j}^T (I - P_A) D_{i,j}^T u = \frac{1}{n} D_{i,j}^T (I - P_A) D_{i,j}^T (I - P_A)^T u .$$

Applying lemmas 7.2 to 7.4 again, shows that these terms converge to a constant, $V_{ij}$ say. Again the convergence is uniform for $\gamma$ in a compact set. The proof is completed by gathering the $V_{ij}$ into a matrix function $V(\gamma)$ of $\gamma$, and observing that $V(\gamma_0) = V_0 . \quad \Box$

7.7 The Operators $C_k C_k^{-1}$

Lemmas 7.2 to 7.4 are proved by construction. In the exponential fitting case, this involves multiplying the discrete Fourier transform of $d_i$ by the transfer functions of $D_u$ and $D_v^T$, and taking the
inverse transform of the result. In preparation for this we need the following lemmas. (Recall that \( \omega \) is the fundamental \( n \)th root of unity.)

**Lemma 7.5.** The sequence

\[ f_k = \rho^{k-1} \quad k = 1, \ldots, n \]

has discrete Fourier transform

\[ F_k = n^{-\frac{1}{2}} (1 - \rho^n) (\omega^{k-1} - \rho^{-1}) \omega^{-k} -1 \quad k = 1, \ldots, n. \]

Proof. Follows from summing a geometric series in \( \omega^{-(k-1)} \rho \), and using \( \omega^n = 1 \).

**Lemma 7.6.** The sequence

\[ F_k = n^{-\frac{1}{2}} (\omega^{k-1} - \rho^{-1}) -2 \omega^{2(k-1)} \quad k = 1, \ldots, n \]

has inverse discrete Fourier transform

\[ f_k = (1 - \rho^n)^{-2} \omega k -1 \quad k = 1, \ldots, n. \]

Proof. Uses geometric series identities, and the fact that

\[ \sum_{j=0}^{n-1} \omega^{mj} = 0 \]

unless \( m = 0 \).

**Lemma 7.7.** If \( p(z) \) is a polynomial of degree less than \( r \), then

\[ F(z) = \frac{p(z)}{(z-a_1) \ldots (z-a_r)} = \sum_{j=1}^{r} \frac{b_j}{z-a_j} \]

with

\[ b_j = Q_j^{-1} p(a_j), \quad Q_j = \prod_{k=1 \atop k \neq j}^{r} (a_j - a_k). \]
Proof. That \( F(z) \) can be expressed as a sum of primitive fractions is the assertion of the partial fraction theorem. The constants \( b_j \) must satisfy

\[
b_j = [(z - a_j)F(z)]_{z=a_j}.
\]

Lemma 7.8. If \( p(z) \) is a polynomial of degree at most \( r \), then

\[
P(z) = \frac{p(z)}{(z - a_1)^2(z - a_2) \cdots (z - a_r)} = \frac{b_1 z}{(z - a_1)^2} + \sum_{j=2}^{r} \frac{b_j}{z - a_j} - \sum_{j=1}^{r} \frac{b_j}{z - a_1}
\]

with

\[
b_1 = \frac{p(a_1)}{a_1 p_1}, \quad b_j = \frac{p(a_j)}{(a_j - a_1) p_j}, \quad q_j = \prod_{k=1, k \neq j}^{r} (a_j - a_k).
\]

Proof. The partial fraction theorem asserts that \( F(z) \) may be written

\[
\frac{b_1 z}{(z - a_1)^2} + \sum_{j=2}^{r} \frac{b_j}{z - a_j} + \frac{b_{r+1}}{z - a_1}.
\]

The constants must satisfy

\[
b_1 = a_1^{-1} [(z - a_1)^2 F(z)]_{z=a_1}
\]

\[
b_j = [(z - a_j)F(z)]_{z=a_j} \quad j = 2, \ldots, r
\]

and

\[
b_1 + b_{r+1} = \left[ \frac{d}{dz} (z - a_1)^2 F(z) \right]_{z=a_1}
\]

\[
= \left[ \frac{d}{dz} \sum_{j=2}^{r} \frac{b_j (a_j - a_1)^2}{z - a_j} \right]_{z=a_1}
\]

\[
= - \sum_{j=2}^{r} b_j.
\]

Proof of lemma 7.2. For rational fitting we have
\[ D_u = (p_{\gamma}(t)^{-1} t^{u-1}) \quad \text{and} \quad a_j = t^{j-1} \]

so we can take
\[ f_j(t) = p_{\gamma}(t)^{-1} t^{u+j-2} . \]

For exponential fitting we have to do much more work. In this case
\[ D_u = \Delta^{T(u-1)} \prod_{j=1}^p (\Delta^T + \zeta_j I)^{-1} \]

which has transfer function
\[ \lambda(z) = \frac{n^{u-1}}{n^p} \frac{(z-1)^{u-1}}{\prod_{j=1}^p (z-\rho_j)} \quad z = \omega, \ldots, \omega^{n-1} . \]

(Recall that \( \rho_j = e^{\frac{j}{n}} \) and \( \zeta_j = n(1-\rho_j) \).) Using lemma 7.5, \( D u_1 \) has discrete Fourier transform
\[ F(z) = \lambda(z) n^{-b_1} \rho_1 z^{l-\rho_1} \frac{1-\rho_n}{z-\rho_1} \quad z = \omega, \ldots, \omega^{n-1} \]

which can be written, using lemma 7.7, as
\[ n^{-b_1} n^{u-1} \rho_1 (1-\rho_1)^{u-1} \left\{ \sum_{j=1}^p \frac{b_j}{z-\rho_j} + \frac{c}{1-z^{1-\rho_1}} \right\} \]

with
\[ b_j = \frac{(\rho_j - 1)^{u-1}}{(1-\rho_j) \prod_{k=1}^p (\rho_j - \rho_k)} \]

and
\[ c = \frac{\rho_1^{-1} - 1)^{u-1}}{\prod_{j=1}^p (\rho_1^{-1} - \rho_j)} . \]

Reversing lemma 7.5, we obtain \( (D u_1)_{\text{s}} \) as
\[
\frac{u-1}{n^p} \rho_1(1-\rho_1^n) \left\{ \sum_{j=1}^{p} \frac{b_j(-\rho_j^{-1})}{1-\rho_j^{-n}} \rho_j^{-s-1} + \frac{c}{1-\rho_1^n} \right\} = \frac{u-1}{n^p} \left\{ - \sum_{j=1}^{p} \frac{b_j \rho_1(1-\rho_1^n)}{1-\rho_j^{-n}} \rho_j^{-s} + c \rho_s \right\}.
\]

Now

\[
\frac{u-1}{n^p} b_j = b_{j,\infty} + O\left(\frac{1}{n}\right), \quad \frac{u-1}{n^p} c = c_{\infty} + O\left(\frac{1}{n}\right)
\]

with

\[
b_{j,\infty} = \frac{(-\beta_j)^{u-1}}{(\beta_1 + \beta_j) \prod_{k=1}^{j} (\beta_k - \beta_j)} \prod_{k=1, k\neq j}^{p} (\beta_1 + \beta_k)
\]

\[
c_{\infty} = \frac{(-\beta_1)^{u-1}}{\prod_{k=1}^{p} (\beta_1 + \beta_k)}.
\]

So let

\[
f_1(t) = c_{\infty} e^{-\beta t} - \sum_{j=1}^{p} b_{j,\infty} \frac{1-e^{-\beta_j t}}{1-e^{-\beta_j}}.
\]

Define \( f_2, \ldots, f_p \) similarly. \( \square \)

Proof of lemma 7.3. The proof is again easy in the rational case since

\[
D_{\gamma u}^{v} = (p_{\gamma}(t)^{u+v-2} t^{u+v-2}).
\]

In the exponential fitting case

\[
D_{\gamma u}^{v} = A^{v-1} A^{T(u-1)} \prod_{j=1}^{p} (A + \epsilon_j I)^{-1} (A^{T} + \epsilon_j I)^{-1}
\]

which has transfer function
\[ \lambda(z) = \frac{n^{u+v-2}}{n^{2p}} \frac{(z-1)^{u-1}(z-1)^{v-1}}{\prod_{j=1}^{p} (z-\rho_j)(z-\rho_j^{-1})} \quad z = \omega^0, \ldots, \omega^{n-1}. \]

Therefore \( D_T^v u_{-1}^* \) has discrete Fourier transform

\[ F(z) = \lambda(z) \frac{n^{-\frac{1}{2}} \rho_1^{1-\rho_1^n} \frac{z}{z-\rho_1}}{\rho_1^{1-\rho_1^n} \prod_{j=2}^{p} (z-\rho_j)(z-\rho_j^{-1})} = \frac{n^{-\frac{1}{2}} \frac{n^{u+v-2}}{n^{2p}} \rho_1^{1-\rho_1^n} \frac{z^{p-u+1}(1-z)^{u-1}(z-1)^{v-1}}{(z-\rho_1)^2 \prod_{j=2}^{p} (z-\rho_j)(z-\rho_j^{-1}) \prod_{j=1}^{p} (1-z\rho_j)}}{\rho_1^{1-\rho_1^n} \prod_{j=2}^{p} (z-\rho_j)(z-\rho_j^{-1})} \]

which can be written, using lemma 7.8, as

\[ n^{-\frac{1}{2}} \frac{n^{u+v-2}}{n^{2p}} \rho_1^{1-\rho_1^n} \frac{z^{b_1 z}}{(z-\rho_1)^2} \sum_{j=2}^{p} \frac{b_j}{(z-\rho_j)^2} + \sum_{j=1}^{p} \frac{c_j}{1-z\rho_j} \frac{1}{(z-\rho_1)} \]

with

\[ b_1 = \rho_1^{(p-u+1)(1-\rho_1^n)} \frac{1}{\rho_1^{1-\rho_1^n} \prod_{k=2}^{p} (\rho_1 - \rho_k) \prod_{k=1}^{p} (1-\rho_k \rho_1)} \]

\[ b_j = \frac{\rho_1^{(p-u+1)(1-\rho_j^n)} \frac{1}{\rho_j^{1-\rho_j^n} \prod_{k=2}^{p} (\rho_j - \rho_k) \prod_{k=1}^{p} (1-\rho_k \rho_j)}}{(\rho_j - \rho_1)^2 \prod_{k=2}^{p} (\rho_j - \rho_k) \prod_{k=1}^{p} (1-\rho_j \rho_k) \prod_{k \neq j}^{p}} \]

and

\[ c_j = \frac{\rho_1^{-(p-u+1)(1-\rho_j^n)} \frac{1}{\rho_j^{1-\rho_j^n} \prod_{k=2}^{p} (\rho_j - \rho_k) \prod_{k=1}^{p} (1-\rho_j \rho_k)}}{(\rho_j - \rho_1)^2 \prod_{k=2}^{p} (\rho_j - \rho_k) \prod_{k=1}^{p} (1-\rho_j \rho_k) \prod_{k \neq j}^{p}} \]

Using lemmas 7.5 and 7.6 to invert \( F(z) \), we obtain \( (D_T^v u_{-1}^*)^* \) as
\[
\frac{n^{u+v-2}}{n^2p} \rho_1(1 - \rho_1^n) \left( \frac{b_1}{(1 - \rho_1^n)^2} s_p^{s-1} + \sum_{j=2}^{P} \frac{b_j}{1 - \rho_j^n} s_j^{s-1} \right) - \sum_{j=1}^{P} c_j \rho_j^{-1} - (s-1) - \sum_{j=1}^{P} \frac{b_j + c_j}{1 - \rho_j^n} \rho_j^{s-1} \right). \\
\]

Note that
\[
\frac{n^{u+v-2}}{n^2p} n_b = b_1^\infty + O(\frac{1}{n}), \quad \frac{n^{u+v-2}}{n^2p} b_j = b_j^\infty + O(\frac{1}{n})
\]
\[
\frac{n^{u+v-2}}{n^2p} c_j = c_j^\infty + O(\frac{1}{n})
\]

with
\[
b_1^\infty = \frac{(-1)^{v-1} n^{u+v-2} \beta_1}{2 \beta_1 \prod_{k=2}^{P} (\beta_k^2 - \beta_1^2)}
\]
\[
b_j^\infty = \frac{(-1)^{v-1} n^{u+v-2} \beta_j}{2 \beta_j (\beta_1 - \beta_j) \prod_{k=1, k \neq j}^{P} (\beta_k^2 - \beta_j^2)}
\]
\[
c_j^\infty = \frac{(-1)^{u-1} n^{u+v-2} \beta_j}{2 \beta_j (\beta_1 + \beta_j) \prod_{k=1, k \neq j}^{P} (\beta_k^2 - \beta_j^2)}
\]

So
\[
g_1(t) = \frac{c_1^\infty}{1 - e^{-\beta_1}} - \frac{\beta_1}{1 - e^{-\beta_1}} te^{\beta_1 t} + \sum_{j=2}^{P} b_j^\infty \frac{1 - e^{-\beta_j}}{1 - e^{-\beta_j}} e^{\beta_j t}
\]
\[
- \sum_{j=1}^{P} c_j^\infty \frac{1 - e^{-\beta_j}}{1 - e^{-\beta_j}} e^{\beta_j t} - \sum_{j=1}^{P} (b_j^\infty + c_j^\infty) e^{\beta_j t}
\]

Define \(g_2, \ldots, g_0\) similarly.
Proof of lemma 7.4. The proof proceeds along very similar lines to that of lemma 7.3, with the difference that all the poles of the discrete Fourier transforms $F(z)$ are simple. Each function $g_{0j}(t)$

$$g_{0j}(t) = e^{-\beta t} \sum_{k=1}^{p} e^{\beta_k t},$$

includes a term in $e^{-\beta t}$ as well as in $e^{\beta_k t}$ and $e^{\beta_k t}$, $k = 1, \ldots, p$. \qed
CHAPTER 8
NUMERICAL EXPERIMENTS

8.1 Programs

The purpose of this chapter is to compare, by way of simulations, the modified Prony algorithm with a good general purpose nonlinear least squares procedure, namely the Levenberg-Marquardt modification of Gauss-Newton. The modified Prony algorithm (henceforth called Prony) was implemented as described in §6.2.2. The symmetric linear system appearing in the inverse iteration sequence was solved by diagonal pivoting, as implemented by J. Bunch in LINPACK (Dongarra et al., 1979). For exponential fitting, the recurrence form was used, and the Prony matrix augmented as described in §6.7.1 to allow for an additive constant in the mean.

The Levenberg-Marquardt algorithm (henceforth called Gauss-Newton) was implemented essentially as described by Osborne (1976). The Levenberg-Marquardt parameter was given an initial value of 1: when necessary it was doubled until a reduction in the sum of squares was achieved; if no increase was required, it was reduced by a factor of 10 for the next iteration.

The convergence criterion used by Gauss-Newton was

\[(8.1) \quad \frac{(ssf^2 - ssr^2)}{(1 + ssf^2)} < \tau\]

where \(ssf\) is the actual sum of squares, and \(ssr\) is the sum of squares based on a linearization of the problem. The tolerance \(\tau\)
was set to $10^{-7}$. The Prony and Gauss-Newton convergence criteria are not strictly comparable. But the Prony tolerance was adjusted, to $10^{-15}$ for exponential fitting and $10^{-10}$ for rational fitting, so that Prony returned estimates that were on average of the same precision as those from Gauss-Newton.

All calculations were performed in double precision in Fortran 77 on a Sperry Univac 1100/82 computer.

8.2 Test Problems

Data was simulated using the mean function

$$\mu(t) = .5 + 2e^{-4t} - 1.5e^{-7t}$$

for exponential fitting, and

$$\mu(t) = (.5 + .5t)/(1 - .5t + .1t^2)$$

for rational fitting. Given a set of 512 standardized random deviates, data sets were constructed to have standard deviations $\sigma = .03, .01, .003, .001$ and sample sizes $n = 32, 64, 128, 256, 512$. The mean functions were evaluated at the equi-spaced points $1/512, \ldots, 1$. Random deviates were associated with the means according to the asymptotic sequence order described in §5.2, so that the data set of size 32 comprises every second point of the data set of size 64 and so on.

For each of the exponential and rational fitting problems, 10 replicates of the data sets were generated. Random deviates were obtained from

$$E = \sigma F^{-1}(U)$$

where $F$ is a standardized distribution function, and $U$ is a pseudo
random number generated by the NAG subroutine G05CAF (Numerical Algorithms Group, 1983) with seed equal to 1984. The actual procedure discarded the first 200 values generated, used the next 5120 for the 10 exponential replications, discarded the next 4680, and used the next 5120 for the rational replications. Four different distribution functions were used: normal, student t on 3 d.f. (infinite third moments), lognormal (skew) and Pareto's distribution with $k = 1$ and $\alpha = 3$ (skew and infinite third moments). But the convergence results were similar for all four distributions, so only those for the normal are reported.

In each problem, the true values themselves were used as starting values.

8.3 Results
8.3.1 Exponential fitting

As table 8.1 shows, Prony requires dramatically fewer iterations than Gauss-Newton for the exponential fitting problem. Furthermore, individual Prony iterations use less machine time on average than those of Gauss-Newton, for which many adjustments of the Levenberg-Marquardt parameter were required. Gauss-Newton was limited to 40 iterations, and was regarded as a failing if it did not converge before this. Prony obliged by always converging, but did so sometimes to complex roots. These were regarded as failures of Prony. Gauss-Newton failed whenever Prony did. For both programs failure occurred when the estimates of $\beta_2$ and $\beta_3$ were relatively close together.

Table 8.2 gives estimated standard deviations averaged over the 10 replications. Reflecting as it does the minimized sums of squares, it gives some idea of comparative precision achieved by the two
Table 8.1. Median and maximum iteration counts, and number of failures, for exponential fitting. Prony above Gauss-Newton.

<table>
<thead>
<tr>
<th>n</th>
<th>σ</th>
<th>.030</th>
<th>.010</th>
<th>.003</th>
<th>.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>6</td>
<td>11</td>
<td>6</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>40</td>
<td>6</td>
<td>33</td>
<td>40</td>
</tr>
<tr>
<td>64</td>
<td>4</td>
<td>8</td>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>32.5</td>
<td>40</td>
<td>5</td>
<td>31.5</td>
<td>40</td>
</tr>
<tr>
<td>128</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>16.5</td>
<td>40</td>
<td>2</td>
<td>10</td>
<td>40</td>
<td>8</td>
</tr>
<tr>
<td>256</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>30</td>
<td>4</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>512</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>36.5</td>
<td>40</td>
<td>5</td>
<td>19.5</td>
<td>40</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 8.2. Mean of $\theta$ over 10 replicates. Given are the leading significant figures, Prony above Gauss-Newton.

<table>
<thead>
<tr>
<th>n</th>
<th>σ</th>
<th>.030</th>
<th>.010</th>
<th>.003</th>
<th>.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>2885</td>
<td>97251</td>
<td>292062</td>
<td>9737559</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2942</td>
<td>98661</td>
<td>293552</td>
<td>9737579</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>2889</td>
<td>96409</td>
<td>289308</td>
<td>9644329</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2914</td>
<td>96879</td>
<td>289484</td>
<td>9644324</td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>2945</td>
<td>98177</td>
<td>294538</td>
<td>9817982</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2950</td>
<td>98259</td>
<td>294538</td>
<td>9818041</td>
<td></td>
</tr>
<tr>
<td>256</td>
<td>2937</td>
<td>97896</td>
<td>293686</td>
<td>9789516</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2940</td>
<td>97925</td>
<td>293686</td>
<td>9789513</td>
<td></td>
</tr>
<tr>
<td>512</td>
<td>2981</td>
<td>99362</td>
<td>298085</td>
<td>9936191</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2983</td>
<td>99376</td>
<td>298085</td>
<td>9936189</td>
<td></td>
</tr>
</tbody>
</table>
algorithms. However the sums of squares are not strictly comparable when complex roots occur. In those cases, Prony always achieves a lower sum of squares by including (implicitly) trigonometric terms in the mean function.

Table 8.3 gives the average estimated rate constants, as calculated by Gauss-Newton. Those for Prony are similar.

Table 8.3. Means and standard deviations of estimates of $\beta_2$ and $\beta_2$ over 10 replications. Gauss-Newton applied to exponential fitting.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\sigma$</th>
<th>.030</th>
<th>.010</th>
<th>.003</th>
<th>.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>.030</td>
<td>4.089 (1.4)</td>
<td>4.127 (.78)</td>
<td>4.138 (.40)</td>
<td>4.065 (.18)</td>
</tr>
<tr>
<td>17.08 (28.)</td>
<td>7.420 (2.0)</td>
<td>6.872 (.82)</td>
<td>6.901 (.36)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>.010</td>
<td>3.937 (1.0)</td>
<td>4.083 (.60)</td>
<td>4.101 (.31)</td>
<td>4.030 (.11)</td>
</tr>
<tr>
<td>8.629 (3.7)</td>
<td>7.169 (1.4)</td>
<td>6.876 (.63)</td>
<td>6.952 (.23)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>.003</td>
<td>3.930 (.66)</td>
<td>4.007 (.39)</td>
<td>4.029 (.20)</td>
<td>4.005 (.06)</td>
</tr>
<tr>
<td>7.680 (1.9)</td>
<td>7.132 (.85)</td>
<td>6.977 (.39)</td>
<td>6.995 (.12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>256</td>
<td>.001</td>
<td>4.022 (.83)</td>
<td>4.071 (.47)</td>
<td>4.024 (.18)</td>
<td>4.004 (.06)</td>
</tr>
<tr>
<td>7.721 (2.1)</td>
<td>7.072 (1.0)</td>
<td>6.992 (.36)</td>
<td>7.001 (.12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>512</td>
<td>.030</td>
<td>4.216 (.65)</td>
<td>4.139 (.37)</td>
<td>4.043 (.13)</td>
<td>4.012 (.04)</td>
</tr>
<tr>
<td>6.974 (1.7)</td>
<td>6.830 (.78)</td>
<td>6.930 (.27)</td>
<td>6.979 (.09)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

8.3.2 Rational fitting

Table 8.4 gives the iteration counts for rational fitting. This problem is much easier for Gauss-Newton, but Prony still takes slightly fewer iterations. For two samples with $n = 32$ and $\sigma = .03$, Prony converged to a stationary value which was not the least squares estimate. With these exceptions, the programs returned effectively
Table 8.4. Median and maximum iteration counts for rational fitting. Prony above Gauss-Newton.

<table>
<thead>
<tr>
<th>n</th>
<th>σ</th>
<th>.030</th>
<th>.010</th>
<th>.003</th>
<th>.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>5</td>
<td>17</td>
<td>6</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>64</td>
<td>5</td>
<td>12</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>8</td>
<td>5.5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>128</td>
<td>6</td>
<td>7</td>
<td>4</td>
<td>2.5</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>256</td>
<td>4.5</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>512</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 8.5. Means and standard deviations, over 10 replicates, of estimates of $\beta_1$ and $\beta_2$. Gauss-Newton applied to rational fitting.

<table>
<thead>
<tr>
<th>n</th>
<th>σ</th>
<th>.030</th>
<th>.010</th>
<th>.003</th>
<th>.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>-0.2910(.24)</td>
<td>-.4474(.055)</td>
<td>-.4853(.015)</td>
<td>-.4952(.005)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.238(.14)</td>
<td>.0692(.034)</td>
<td>.0914(.095)</td>
<td>.0972(.003)</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>-.5042(.21)</td>
<td>-.5025(.066)</td>
<td>-.5009(.020)</td>
<td>-.5003(.007)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.1033(.12)</td>
<td>.1017(.038)</td>
<td>.1006(.011)</td>
<td>.1002(.004)</td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>-.4694(.14)</td>
<td>-.4918(.046)</td>
<td>-.4977(.013)</td>
<td>-.4993(.005)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.0816(.08)</td>
<td>.0950(.027)</td>
<td>.0986(.008)</td>
<td>.9954(.003)</td>
<td></td>
</tr>
<tr>
<td>256</td>
<td>-.4979(.13)</td>
<td>-.5002(.041)</td>
<td>-.5002(.012)</td>
<td>-.5001(.004)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.0966(.08)</td>
<td>.0994(.025)</td>
<td>.0999(.007)</td>
<td>.1000(.002)</td>
<td></td>
</tr>
<tr>
<td>512</td>
<td>-.5180(.16)</td>
<td>-.5080(.051)</td>
<td>-.5026(.015)</td>
<td>-.5009(.005)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.1097(.09)</td>
<td>.1044(.030)</td>
<td>.1014(.009)</td>
<td>.1005(.003)</td>
<td></td>
</tr>
</tbody>
</table>
identical estimates. Table 8.5 gives the average estimates for $\beta_1$ and $\beta_2$, as calculated by Gauss-Newton.

8.4 Discussion

8.4.1 Prony convergence criterion

The eigenstructure of $B_c$, given by theorem 6.3, raises a problem for the convergence criterion for the exponential fitting problem. In our simulations we have $p = 3$, so the largest eigenvalue of $n^{-1}B_c$ is $O(n^6)$. For $n = 512$, this effect is large enough to explain the one step convergence by itself. The Prony algorithm returns very creditable estimates though, despite this, and a fully satisfactory explanation of this phenomenon is yet to be made. An attempt to sidestep the problem, by calculating the difference form Prony matrix using (5.19) and (5.14), was defeated by subtractive cancellation. Resolution of this problem waits upon a direct implementation of the difference form algorithm.

8.4.2 Unmodified Gauss-Newton

Two changes to the Levenberg-Marquardt algorithm used allow it to converge rather more rapidly on the rational fitting problem. The first is to remove the constant from the denominator of the convergence criterion (8.1). The number of iterations required then decreases consistently with $n$, as would be expected from curvature considerations. The second is to set the Levenberg-Marquardt parameter to zero. The message is that rational fitting is not a highly nonlinear problem, so that the unmodified Gauss-Newton algorithm performs very well.
REFERENCES


Wald, A. (1943). 'Tests of statistical hypotheses concerning several parameters when the number of observations is large', Trans. Amer. Math. Soc. 54:426-482.


